analytical chemistry

Article

pubs.acs.org/ac

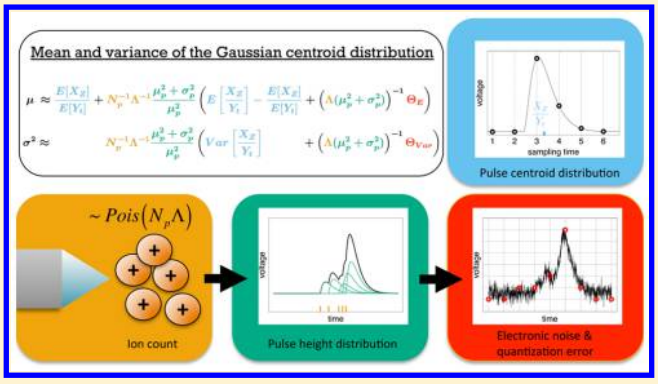
Derivation of the Statistical Distribution of the Mass Peak Centroids of Mass Spectrometers Employing Analog-to-Digital Converters and Electron Multipliers

Andreas Ipsen*

Institute of Mass Spectrometry, College of Medicine, Swansea University, Swansea, Wales SA2 8PP, United Kingdom
Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352, United States

Supporting Information

ABSTRACT: The mass peak centroid is a quantity that is at the core of mass spectrometry (MS). However, despite its central status in the field, models of its statistical distribution are often chosen quite arbitrarily and without attempts at establishing a proper theoretical justification for their use. Recent work has demonstrated that for mass spectrometers employing analog-to-digital converters (ADCs) and electron multipliers, the statistical distribution of the mass peak intensity can be described via a relatively simple model derived essentially from first principles. Building on this result, the following article derives the corresponding statistical distribution for the mass peak centroids of such instruments. It is found that for increasing signal strength, the centroid distribution converges to a Gaussian distribution whose mean and variance are determined by physically meaningful parameters and which in turn determine bias and variability of the m/z measurements of the instrument. Through the introduction of the concept of "pulse-peak correlation", the model also elucidates the complicated relationship between the shape of the voltage pulses produced by the preamplifier and the mean and variance of the centroid distribution. The predictions of the model are validated with empirical data and with Monte Carlo simulations.



The development of new methods of MS data analysis is a very active area of research, which often involves the use of "error models" that describe the measurement error associated with the measurement of the mass peak intensity and centroid. However, while the applications of such models are widely disseminated in the literature the more foundational task of establishing the nature of the "true" error model, a task whose resolution is in many ways a prerequisite for addressing downstream analyses in a statistically defensible manner, has received relatively little attention. This disparity of research efforts is particularly marked when leaving aside studies that attempt to establish error models using heavily preprocessed data or which lack detailed consideration of the instrumental operations, either of which effectively preclude the models obtained from describing the distribution of the data dependably.

An unavoidable feature of any model that attempts to describe MS data from first principles is that it will only be applicable to mass spectrometers with the particular instrumental configuration considered in the model's derivation. The model that will be derived in the present study will therefore only apply to instruments that use ADCs and electron multipliers, which includes most modern time-of-flight (TOF) instruments, as well as many sector and quadrupole instruments. The model will also apply to instruments that are not mass spectrometers but which do use such a setup.

Among the relatively small number of studies that investigate the mass peak centroids of such instruments from instrumental fundamentals an application note by Gedcke[1] provides a particularly detailed discussion of the statistical aspects of the ADC digitization process, although it ultimately does not provide a fully rigorous mathematical treatment of it. A number of other studies[2−6] place the emphasis on determining the distribution of ion arrivals while providing a relatively brief account of the effects of the detector and subsequent electronics. Two further studies by Harris et al.[7] and Peterson and Hayes[8] provide very detailed discussions of the mass peak intensity distribution but without extending their models to cover the centroid distribution and again without a fully rigorous accounting of the effects of the digitization process. It should be noted that the literature on the alternative but currently less widely used, time-to-digital (TDC) digitization system is somewhat more developed.[9−13]

Received: June 26, 2016
Accepted: January 13, 2017
Published: February 3, 2017

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

The study presented here provides a detailed derivation from first principles of the statistical distribution of mass peak centroids for ADC-based mass spectrometers employing electron multipliers. As will be demonstrated, where instrumental parameters are known exactly, the distribution of mass peak centroids is approximately Gaussian, provided the ion count and the signal-to-noise ratio of the mass peak intensity are not too low. The mean and variance of this Gaussian can be linked directly to a range of fundamental parameters, including the mean number of ion arrivals at the detector, the mean and variance of the electron multiplier's pulse height distribution,[14] and to the shape of the voltage pulse produced by the preamplifier. It is also linked to the idealized centroid distribution produced by a single such normalized voltage pulse in the absence of electronic noise and quantization noise. Where instrumental parameters must be estimated through the use of reference compounds, so that they are subject to statistical uncertainty, the distribution will generally be more complicated than a Gaussian, but depending on the calibration procedure used, it might ultimately still be determined using the results derived here.

The present paper makes extensive use of the results of a previous study on the statistical distribution of the mass peak intensity.[15] However, the modeling of the mass peak centroid will prove to be more complicated in the sense that more demanding algebra is required, and somewhat more diverse mathematical methods are needed. While the mass spectrometry community has long accepted the need for very intricate mathematics in ion optical modeling,[16−19] there is not much tradition of comparably detailed mathematical efforts in the statistical modeling of the digitization process. The findings of this study suggest that this will have to change if a full understanding of MS data is to be attained.

THEORY

Derivation Outline. The centroid is the weighted mean of the sampling times across a mass peak, where the weights are given by the relative intensities associated with the respective sampling times. More formally, suppose a mass peak is centroided across a set of M_p sampling times, each separated by a time period Δt, the first one being at time t_0 and the last being at time $t_0 + \Delta t(M_p - 1)$. If y_i is the intensity recorded at the ith sampling time then the centroid may be defined as

$$C = \frac{\sum_{i=0}^{M_p-1} (t_0 + i\Delta t) y_i}{\sum_{i=0}^{M_p-1} y_i} = \frac{J}{H} \tag{1}$$

Consequently, the centroid can be regarded as the division of the random variable J, which is effectively the inner product of the intensities observed across a mass peak and the associated sampling times, by the random variable H, which is the sum of the intensities observed across the mass peak. Therefore, the distribution of C is that of the ratio of two correlated random variables. While a ratio distribution is a complicated quantity, its mean and variance can be approximated via Taylor expansions[20] as

$$E[C] = E\left[\frac{J}{H}\right] \approx \frac{E[J]}{E[H]} - \frac{\mathrm{Cov}[H, J]}{E[H]^2} + \frac{E[J]}{E[H]^3}\mathrm{Var}[H] \tag{2}$$

and

$$\mathrm{Var}[C] = \mathrm{Var}\left[\frac{J}{H}\right] \approx \frac{\mathrm{Var}[J]}{E[H]^2} - \frac{2E[J]}{E[H]^3}\mathrm{Cov}[H, J] + \frac{E[J]^2}{E[H]^4}\mathrm{Var}[H] \tag{3}$$

where the first and second order terms are included for the mean, but only the first order terms are included for the variance, in order to keep the expression manageable. While this only accounts for the first two moments of the distribution of C, it will very often suffice in practice, since as will be illustrated, C quickly converges to the Gaussian distribution for increasing signal strength. This means that the centroid will very often be Gaussian, regardless of the shape of the digitized mass peak, regardless of the distribution of the ion arrival times at the electron multiplier (hence regardless of the type of mass analyzer used) and regardless of the pulse height distribution. This useful result may be regarded as a consequence of the Delta method[20,21] since both H and J converge to the Gaussian distribution by the arguments given in earlier work.[15] Thus, it will be possible to approximate the distribution of C very closely, by determining the means, variances, and covariance of H and J.

System Summary. The physical processes that govern the generation of MS data have been discussed in detail in earlier work[15] and are summarized with simulated data on Figure 1. An overview of the variables and parameters used in modeling these processes is provided in Table S-1 of the Supporting Information. The total number of ions striking the instrument's electron multiplier over the duration of a mass peak will be an outcome of the Poisson distribution with mean Λ. When a single one of these ions strikes the electron multiplier, a current pulse is produced which is passed to a preamplifier which amplifies the signal and turns it into a voltage pulse. While the normalized shape of this voltage pulse is given by the function $f(t)$, its area is a random variable that has mean μ_p and variance σ_p^2 and which is determined by the electron multiplier, by the properties of the ion striking it, and by the preamplifier's gain. Such a voltage pulse will be produced in response to each of the ions striking the electron multiplier and where these pulses overlap they are superposed.

Electronic noise is introduced in the form of white noise with variance σ_s^2 and noise determining the baseline, with variance σ_B^2 and mean μ_B where, in the present study, it will be assumed that the ADC quantization levels are chosen such that $\mu_B = 0$. The resulting superposition of voltage pulses and electronic noise is assumed to be within the linear range of the electronics and is digitized by an ADC with time resolution Δt and voltage resolution ΔV. In order for the present model to apply the ADC should be unsaturated at both extremes, which is an assumption that is typically not satisfied by commercial instruments, which often suppress the appearance of electronic noise by setting the lowest ADC reference voltage above the noise floor. Additional dither noise may be deliberately introduced in order to render the quantization error mathematically tractable, resulting in a quantization error with variance $d\Delta V^2$ where the value of d depends on the nature of the dither used.[22]

The digitized intensities of N_p distinct acquisitions are summed or "histogrammed" in memory to form a single spectrum, which is then written to disk. The intensity, H, of a mass peak in this spectrum is obtained by summing the

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

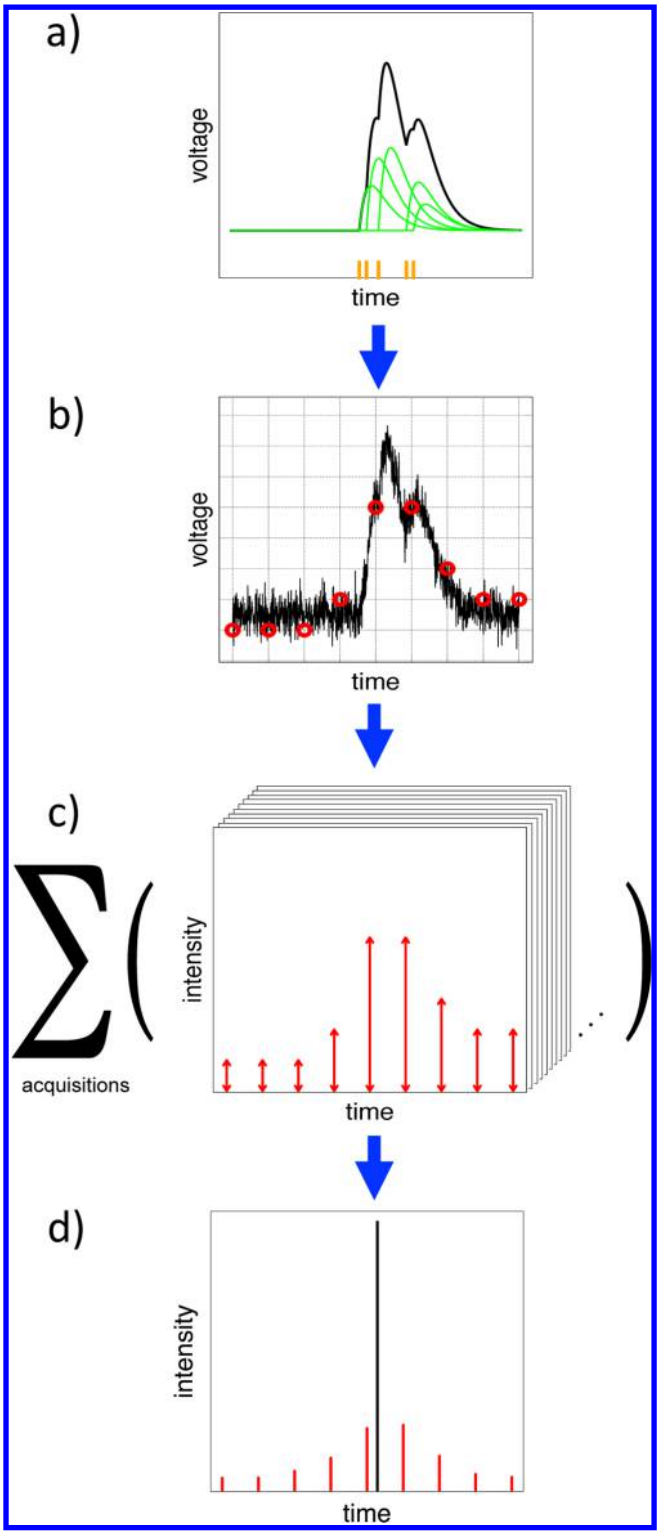


Figure 1. (a) Five ions strike the detector at distinct times (orange) over the duration of a mass peak, causing the preamplifier to produce five voltage pulses (green) whose superposition (black) is to be digitized by the ADC. (b) The preamplifier also introduces electronic noise which is superposed onto the signal, and the resulting noisy signal (black) is sampled every Δt seconds by the ADC (as indicated by the vertical gridlines) and rounded to a set of discrete levels (indicated by the horizontal gridlines) which are determined by the ADC's reference voltages. The signal therefore gets digitized to the red circles, which are stored in memory. (c) Multiple acquisitions (often hundreds) are recorded to memory in this way and the intensities of matching sampling times are then summed or "histogrammed" to produce a mass peak in "profile mode" as indicated by the red spectrum in panel d which may be written to disk. In "centroid mode" the mass peak is further reduced to the black bar which is characterized by a single intensity, H, (the sum of all intensities recorded across the mass peak profile), and a centroid, C (the weighted average of the sampling times across the mass peak with the weights given by the histogrammed intensities). Many commercial instruments set the lowest reference voltage of the ADC above the noise floor so that the intensities at the edges of a mass peak profile appear fixed at zero. Figures adapted from ref 15. Copyright 2015 American Chemical Society.

recorded intensities across the M_p sampling times that fully span that mass peak. The random variable J is obtained by first multiplying the intensities by the associated sampling times and then summing them, and the mass peak's centroid, C, is the division of J by H. The average of the sampling times over which the sum is taken will be labeled t_M. It is assumed that distinct electronic noise terms are independent, that Λ is large enough for the Central Limit Theorem to apply, and that the standard deviation of the distribution of ion arrival times is greater than Δt. The model that will be developed is applicable to the joint centroid induced by overlapping mass peaks if those peaks are induced by ions with the same pulse height distribution, as would be the case, for example, for the unresolved mass peaks of the isotopic fine structure. The mass window over which the centroid is calculated should not overlap with any outside mass peaks.

Idealized Sampling. In analyzing the distributions of both H and J it is useful to take the same approach used in earlier work[15] and consider first a simplified system that will be referred to as "idealized sampling" wherein (1) all voltage pulses produced 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. The distributions of both H and J can be determined by gradually removing the above assumption. In section 2 of the Supporting Information, the means, variances, and covariance of H and J are obtained when multiple ion arrivals and variable pulse heights are allowed. In section 3 of the Supporting Information, electronic noise and finite voltage resolution are accounted for through the introduction of Y_F and X_F, which are the analogues of H and J when the signal being digitized from a single acquisition is solely due to electronic noise. In section 4 of the Supporting Information, histogramming is also accounted for, resulting in the full, nonidealized expressions for the means, variances, and covariance of H and J, which are used to determine the mean and variance of C in sections 5 and 6 of the Supporting Information, respectively. The distribution of C can be understood without a detailed examination of its derivation. However, in order to express it a number of conceptually important variables must first be defined under the assumption of idealized sampling.

We define the random variable Z with mean μ_Z and (finite) variance σ_Z^2, to be the arrival time of the voltage pulse induced by a single ion at the ADC. Z generally approximates the arrival times of the ions at the detector and is therefore primarily determined by the type of mass analyzer used; however, it is offset by the time it takes the resulting signal to reach the ADC.

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

It may also be affected by other distortions such as the relationship between the ion's impact site on the electron multiplier and the signal's transit time through it.[3] Z will typically be a mixture distribution that arises from the unresolved fine-structure isotope pattern that induces the mass peak under analysis. It is useful to write Z as

$$Z = \lfloor Z \rfloor + V \tag{4}$$

where the first term rounds the voltage pulse arrival time down to the time of the nearest preceding sampling time, while the second term accounts for the time interval between that sampling time and the actual pulse arrival time.

We define Y_t to be the sum of the intensities observed over the M_p sampling times in response to a single normalized voltage pulse:

$$Y_t = \sum_{i=0}^{M_p-1} f(t_0 + i\Delta t - Z) \tag{5}$$

and for the idealized system considered here $H = Y_t$.

Similarly, for this idealized system J is equal to a quantity that will be labeled X_Z. It is given by the sum of the M_p sampling times (t_0, $t_0 + \Delta t$, ..., $t_0 + (M_p - 1)\Delta t$), each multiplied by the value of the voltage pulse at that sampling time ($f(t_0 - Z)$, $f(t_0 + \Delta t - Z)$, ..., $f(t_0 + (M_p - 1)\Delta t - Z)$):

$$X_Z = \sum_{i=0}^{M_p-1} (t_0 + i\Delta t) f(t_0 + i\Delta t - Z) \tag{6}$$

As demonstrated in section 1 of the Supporting Information, X_Z can be rewritten as

$$X_Z = \lfloor Z \rfloor Y_t + X_t \tag{7}$$

where $\lfloor Z \rfloor Y_t$ reflects the contribution to X_Z determined by the number of whole sampling intervals that precede the voltage pulse arrival, while X_t reflects the contribution determined by the pulse's precise arrival time within a sampling interval and is defined as

$$X_t = \sum_{i=0}^{M_p-1} i\Delta t f(i\Delta t - V) = \sum_{i=0}^{M_p-1} g(i\Delta t - V) \tag{8}$$

where the function $g(t)$ is defined as

$$g(t) = \Delta t \left\lfloor \frac{t}{\Delta t} \right\rfloor f(t) \tag{9}$$

so that $g(t) = \Delta t f(t)$ for $t \in [0, \Delta t]$, $g(t) = 2\Delta t f(t)$ for $t \in (\Delta t, 2\Delta t]$, $g(t) = 3\Delta t f(t)$ for $t \in (2\Delta t, 3\Delta t]$, etc., as illustrated in Figure S-1 of the Supporting Information, where example instances of $f(t)$ and $g(t)$ are shown. While the total area under $f(t)$ is by definition 1, the total area under $g(t)$, will be labeled G.

The distribution of C under idealized sampling will be referred to as the *pulse centroid distribution* and as will be demonstrated; its mean and variance shape the mean and variance of C in a nonidealized system. It may be written

$$\frac{X_Z}{Y_t} = \frac{\sum_{i=0}^{M_p-1} (t_0 + i\Delta t) f(t_0 + i\Delta t - Z)}{\sum_{i=0}^{M_p-1} f(t_0 + i\Delta t - Z)} \tag{10}$$

It will be assumed that Δt is small enough, relative to $f(t)$, that eqs 2 and 3 can be used to approximate the mean and variance of the pulse centroid distribution. This is a fair assumption since an instrument for which Δt were so large that $f(t)$ could effectively "fall between" sampling times would generate mass peak intensities with very poor signal-to-noise ratios.

Mean and Variance of the Mass Peak Centroid. The mean and variance of the mass peak centroid, C, can be obtained from eqs 2 and 3 as shown in sections 5 and 6 of the Supporting Information. Both the mean and the variance of C can be decomposed into two sets of terms—one relating to the pulse centroid distribution and another that relates to the electronic noise and quantization errors.

The mean of the centroid can be written:

$$E[C] \approx \frac{E[X_Z]}{E[Y_t]} + N_p^{-1}\Lambda^{-1}\frac{\mu_p^2 + \sigma_p^2}{\mu_p^2}\left(E\left[\frac{X_Z}{Y_t}\right] - \frac{E[X_Z]}{E[Y_t]} + (\Lambda(\mu_p^2 + \sigma_p^2))^{-1}\Theta_E\right) \tag{11}$$

The ratio $E[X_Z]/E[Y_t]$ is also the first order approximation of the mean of the pulse centroid distribution. It can be written explicitly as

$$\frac{E[X_Z]}{E[Y_t]} = \mu_Z + G - \Delta t/2 + \Delta t\ \mathrm{Cov}[\lfloor Z \rfloor, Y_t] \tag{12}$$

Its first term, μ_Z, is the mean pulse arrival time, and this is generally the only parameter of the mean of C that is of interest to the analyst; the remaining terms constitute various forms of bias that should be corrected for. $G - \Delta t/2$ accounts for the contribution due to the finite duration of the voltage pulse in the absence of *pulse-peak correlation*, which is a concept that will be discussed in the next subsection and which is accounted for by the final covariance term.

It is noteworthy that the higher order terms of the mean tend to zero with increasing N_p and Λ. For mass peaks with a high total ion count, these terms may therefore be small enough that correcting for them is unnecessary. Moreover, they can be reduced by reducing the coefficient of variation of the pulse height distribution, σ_p/μ_p. The quantity $E[X_Z/Y_t] - E[X_Z]/E[Y_t]$ accounts for higher order terms of the mean of the pulse centroid distribution. These are written out explicitly in the Supporting Information, but they vanish in the absence of pulse-peak correlation as described in the next subsection. The quantity Θ_E accounts for the bias that is due to the electronic noise and quantization errors. It can be written more explicitly as

$$\Theta_E = E[Y_t]^{-2}\ \mathrm{Var}[Y_F]\left(\frac{E[X_Z]}{E[Y_t]} - t_M\right) \tag{13}$$

The closer the ADC sampling times are centered on the first order mean of the centroid, the closer this component of the bias will be to zero. Furthermore, its magnitude relative to the other second order terms decreases with the second raw moment of the pulse height distribution and with increasing ion count.

The variance of C can be written

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

$$\mathrm{Var}[C] \approx N_p^{-1}\Lambda^{-1}\frac{\mu_p^2 + \sigma_p^2}{\mu_p^2}\left(\mathrm{Var}\left[\frac{X_Z}{Y_t}\right] + (\Lambda(\mu_p^2 + \sigma_p^2))^{-1}\Theta_{\mathrm{Var}}\right) \quad (14)$$

where the same terms that scale the second order terms of the mean are seen to scale the variance. This highlights the importance to good mass accuracy of high ion counts and a pulse height distribution with a low coefficient of variation. The importance of the former is widely accepted,[23−25] but the latter factor appears to have received relatively little attention in the context of mass accuracy.

$\mathrm{Var}[X_Z/Y_t]$ is the variance of the pulse centroid distribution, and as shown in section 6 of the Supporting Information, for $\sigma_Z > \Delta t$ it can be written out more explicitly as

$$\mathrm{Var}\left[\frac{X_Z}{Y_t}\right] \approx \Delta t\alpha(Gf - g, \Delta t) + 2\,\mathrm{Cov}\left[\frac{X_t}{Y_t}, \lfloor Z \rfloor\right] + \sigma_Z^2 + \frac{1}{12}\Delta t^2 \quad (15)$$

where $\alpha(Gf - g, \Delta t)$ is the sum of the autocorrelation of the function $Gf - g$ over all lags that are integer multiples of Δt. The covariance term is a further feature of the pulse-peak correlation, and it is apparent that a negative correlation of the terms involved can in this case play a role in reducing the centroid's variance. σ_Z^2 is the variance of the pulse arrival time, Z, whose minimization is the conventional focus of ion optics, while the $\Delta t^2/12$ term results from the time discretization that is inherent to the digitization process.

Θ_{Var} can be written out more explicitly as

$$\Theta_{\mathrm{Var}} = E[Y_t]^{-2}\,\mathrm{Var}[Y_F]\left(\frac{E[X_Z]}{E[Y_t]} - t_M\right)^2 + E[Y_t]^{-4}(\sigma_s^2 + d\Delta V^2)\left(\frac{M_p^3 - M_p}{12}\right) \quad (16)$$

where, again, the importance of centering the sampling times on the first order mean of the centroid is evident from the first term, which vanishes when this centering is exact. The importance of using a small set of sampling times (while leaving out only a negligible portion of the signal induced by the incoming ions) is highlighted by the cubed M_p in the second term.

Pulse-Peak Correlation. A large part of the reason why the algebra required to obtain the above expressions is so complicated is that Y_t (the idealized intensity of a single voltage pulse) generally varies with the random variable V, which determines the precise voltage pulse arrival time within a sampling interval as illustrated on Figure 2. Since the conditional distribution of V can vary considerably across the different sampling intervals spanning a mass peak, the portions of $f(t)$ that are likely to get sampled will also vary across the mass peak and this is what gives rise to pulse-peak correlation.

The effects of pulse-peak correlation can be made somewhat more tangible by generating simulated centroids for the two pulse shapes shown on Figure 2 and comparing the centroid distributions obtained for a Gaussian Z with those obtained when Z has a uniform distribution that spans an integer number of sampling intervals. The latter distribution has the special

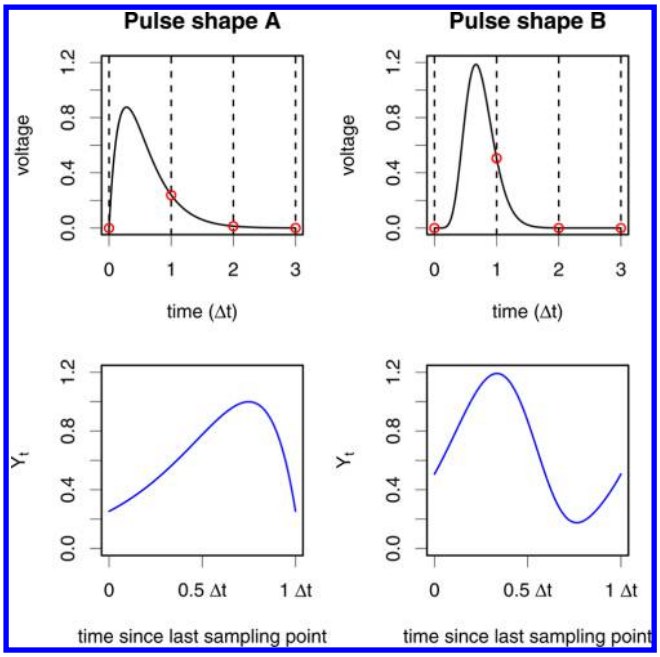


Figure 2. Comparison of the intensities induced by two different normalized pulse shapes that have the same value of G. The top plots show the two pulse shapes when their arrival times are at time 0, and the red circles indicate the values that are sampled and summed to produce Y_t. The bottom plots show the respective intensities as a function of the time elapsed since the latest sampling time preceding the pulse arrival time (the random variable V). Clearly pulse shape A will tend to induce a greater intensity if it arrives in the later half of a sampling interval, whereas pulse shape B will tend to induce a greater intensity if it arrives in the earlier half. Since ions are most likely to arrive late in a sampling interval on the low-mass side of a Gaussian mass peak and early on the high-mass side, pulse shape A will on average "drag down" the centroid, whereas pulse shape B will "push it up".

property that V will have the same distribution across all sampling intervals, so that there is no pulse-peak correlation in eq 12. On Figure 3, 1000 centroids are simulated for $N_p\Lambda = 1000$ and without electronic noise, so that a first order approximation of the centroid means is adequate. For the Gaussian Z, this results in two clearly distinct populations, whose means are displaced from $\mu_Z + G - \Delta t/2$ by the distance predicted by the covariance term in eq 12. However, simulating centroids under conditions that are exactly identical, except for the use of a uniform Z spanning exactly four sampling intervals (but with the same mean and variance as the Gaussian Z) results in both centroid distributions being centered on $\mu_Z + G - \Delta t/2$. This is because for the uniform Z, the Y_t profiles illustrated on Figure 2 are sampled from uniformly across all sampling intervals. Additional parameters used for the simulations, which as with all centroid simulations used in this study were run with a custom written R script, are listed in Table S-2 of the Supporting Information.

In practice it is not feasible to produce uniform pulse arrival distributions that span an integer number of sampling intervals. Moreover, the pulse arrival distribution will generally vary with factors such as mass and charge state, as will its alignment with the ADC sampling times. Therefore, the bias to the centroid mean that is due to the pulse-peak correlation will potentially be difficult to correct for, as it may affect reference compounds used as part of mass calibration procedures differently than it

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

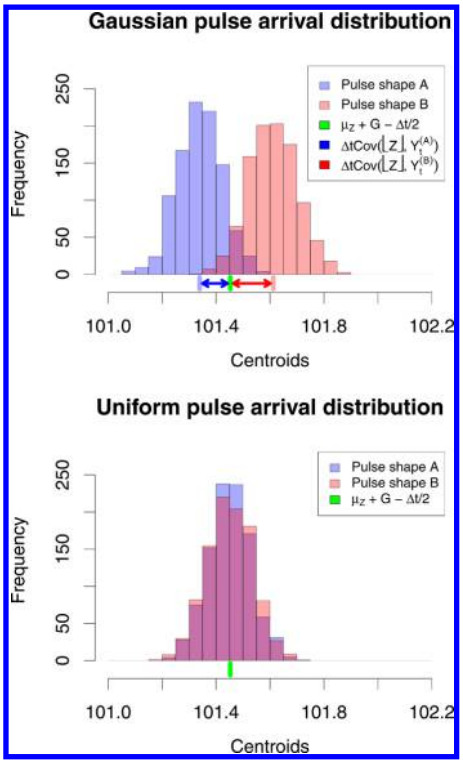


Figure 3. Comparison of the distributions of centroids obtained from the two pulse shapes shown on Figure 2 when the pulse arrival distribution is Gaussian (top) and uniform (bottom). For the former, the pulse shapes are sampled differently across the different sampling intervals causing the means of the two populations of centroids to be displaced from one another as predicted by the pulse-peak correlation.

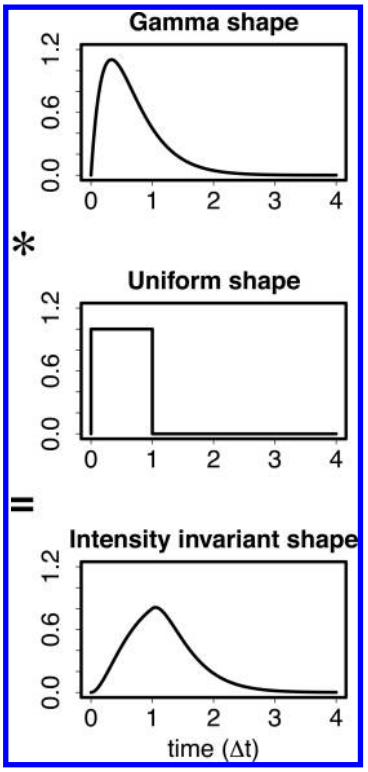


Figure 4. (Top) Probability density function of the gamma distribution with rate parameter $3\Delta t$ and shape parameter $2\Delta t$. (Middle) Normalized rectangular function ranging from 0 to Δt. (Bottom) Intensity invariant voltage pulse obtained through the convolution of the above functions.

affects analytes of interest, and unlike the higher order biases, its effects do not diminish with increasing ion counts.

These considerations would be significantly simplified if the value of Y_t were constant across all sampling intervals. The variability of Y_t can be reduced by using faster (and more expensive) digitizers or alternatively $f(t)$ might be made broader relative to Δt, although this would necessitate increasing M_p which increases the centroid variance. However, it is in principle possible to obtain entirely constant Y_t with short but appropriately shaped pulses. For example a rectangular voltage pulse of duration Δt, will result in a constant Y_t and thereby maximize the signal-to-noise ratio of the mass peak intensity.[15] A wide range of pulse shapes can be produced through the use of appropriate pulse shaping circuits, but it is not easy to produce an approximately rectangular pulse for fast digitizers for which Δt can be less than one nanosecond. However, the class of *intensity invariant voltage pulses*, defined as those for which Y_t is constant, is much broader. In fact as demonstrated in the Supporting Information (section 7) intensity invariant voltage pulses may be generated by taking the convolution of any properly normalized pulse shape with the normalized rectangular function of length Δt. An example of such a pulse, obtained through the convolution of the rectangular function with the probability density function of the gamma distribution, is provided in Figure 4. In addition to maximizing the signal-to-noise ratio of the mass peak intensity such pulses completely eliminate the m/z bias due to the pulse-peak correlation, irrespective of the ion count, which greatly simplifies downstream analyses. Although the design and analysis of pulse shaping circuits is a well-developed area of nuclear electronics,[26] it is not clear how thoroughly it has been explored in the context of mass spectrometry.

EXPERIMENTAL SECTION

The above model was validated using a subset of the data set used previously,[15] which was acquired with a MAT95 sector mass spectrometer (Finnigan MAT GmbH, Bremen, Germany) that incorporates a discrete dynode multiplier (MasCom Multiplier model MC 17A) for signal detection and whose preamplifier (Thermo Finnigan part no. 2064460 incorporating a Texas Instruments OPA602A operational amplifier) was monitored with a 12-bit 1 GHz LeCroy HD04000 oscilloscope connected to one of the preamplifier's test points. The voltage pulses produced by this preamplifier are considerably longer than those that would be produced by a typical TOF instrument but are also sampled at a much lower frequency. The mass spectrometer was operated in electron ionization positive ion mode and ^{36}Ar ions from residual air were used in the analysis due to the low degree of interference of their mass peaks from distinct ion species. The MAT95 was set to continually scan over the appropriate mass region while the oscilloscope was set to trigger on the signal from the resulting mass peak, as it proved impractical to use the mass spectrometer's start signal. A total of 489 mass peaks were ultimately recorded and the functional form of $f(t)$ was identified as being extremely close to the probability density function of the gamma distribution (Figure S-2 of the Supporting Information). This functional form was fitted to the signals induced by all incoming ions via a custom written R script, so that their arrival times and pulse heights could be

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

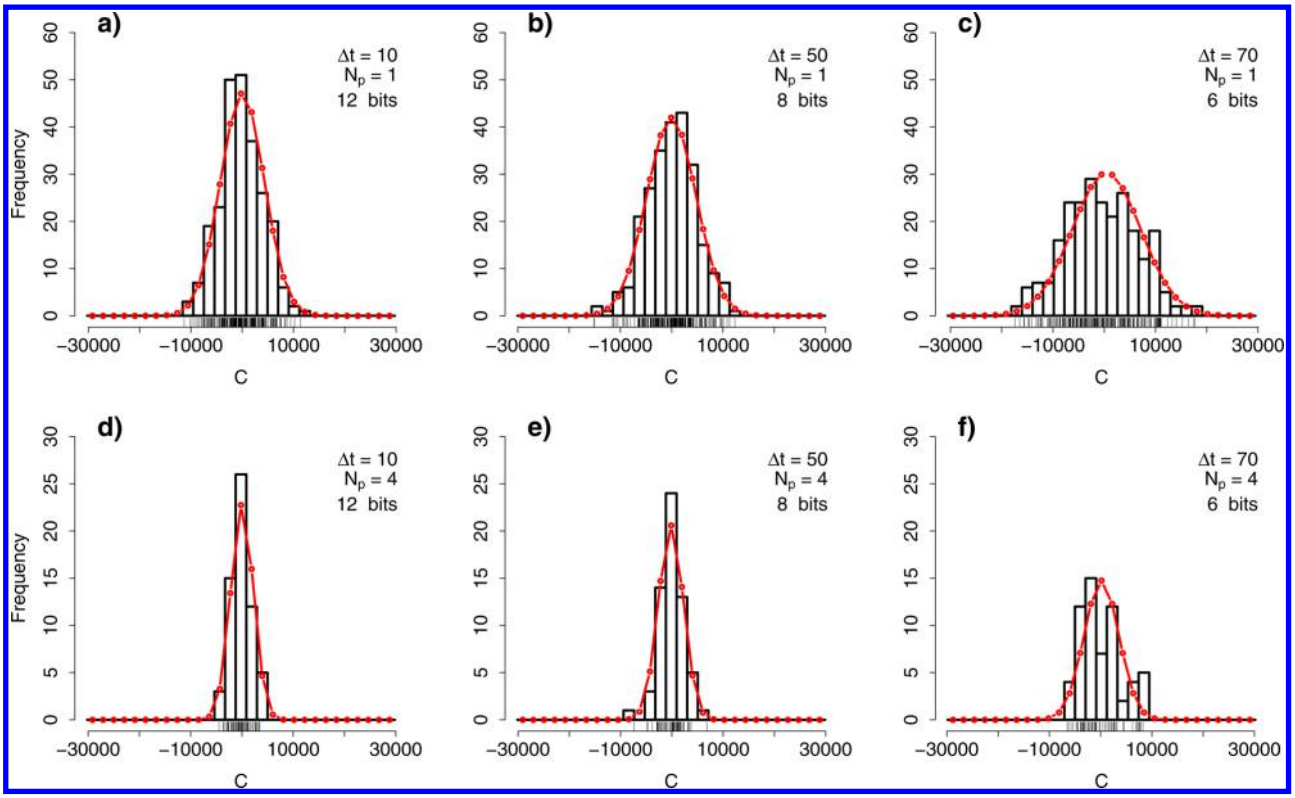


Figure 5. Comparison of the Gaussian distributions predicted by eqs 11 and 14 (red) with the empirical distributions of the mass peak centroids (black) for the six instrumental settings listed in the top right corners. Panels a, b, and c on the top row all show the centroids of mass peaks obtained from single acquisitions but with decreasing voltage and time resolution going from left to right. Panels d, e, and f on the bottom row show the centroids of mass peaks obtained by histogramming four acquisitions, again with decreasing voltage and time resolution going from left to right.

determined with a very high degree of accuracy. The data were standardized so that $\Delta t = 1$ and $\Delta V = 1$.

As previously noted,[15] there were clear signs of electromagnetic interference in the data, whose origin could not be established, and this caused significant distortions to the distribution of C. Consequently the interference was filtered out by fitting the known functional form of $f(t)$ to all observed signals and working with the fitted values (see Figure S-2 of the Supporting Information). This also had the effect of filtering out the high-frequency noise, which was therefore reintroduced by adding high-frequency Gaussian noise with the same variance as the original high-frequency noise at each sampling time. The intensities assigned to the ADC quantization levels were chosen such that μ_B was zero, as was assumed in the derivation of the model.

Because it was not possible to trigger the oscilloscope with the start signal of the mass spectrometer, the experimental setup did not allow for the simplifying assumption that the distributions of ion arrival times of different acquisitions have the exact same means. However, since the arrival times of effectively all ions had been determined by fitting $f(t)$ to the data, the 489 spectra could be realigned such that the distribution of sample means of the ion arrivals differed only by an amount consistent with their natural statistical variation. The time reference was chosen such that the underlying ion arrival mean, μ_Z, was zero.

As discussed previously,[15] a wide range of experimental settings may be replicated by applying appropriate operations to the data. Specifically the digitizer speed may be reduced to $\Delta t = 2$, $\Delta t = 3$, ..., by sampling only every 1 in 2, 1 in 3, etc., time steps, while the voltage resolution can be reduced below 12 bits by rounding the intensities appropriately, and the value of N_p can be increased by histogramming groups of spectra. A total of 12 different combinations of instrumental settings (listed in Figure 5 and in Figure S-3 of the Supporting Information) were considered in this manner.

Following the previously used protocol,[15] half of the spectra were used to calculate the values of instrumental parameters, such as the rate of ion arrivals, Λ, the mean and variance of the pulse height distribution, μ_p and σ_p^2, and the variances of the electronic noise, σ_s^2 and σ_B^2. The terms of the pulse-peak correlation were calculated from the sample covariances of 10^5 simulated voltage pulses having the same functional form as the empirically observed $f(t)$ and simulated under idealized sampling. On the basis of these and the other known parameters (listed in Table S-2 of the Supporting Information), the mean and variance of C as predicted by eqs 11 and 14 could be calculated, and the corresponding Gaussian distribution compared to the empirical distribution of C obtained from the remaining half of the spectra. This was done for each of the 12 sets of instrumental settings considered as illustrated in Figure 5 and Figure S-3 of the Supporting Information. Quantile-quantile plots comparing the predicted and empirical distributions are also provided on Figures S-4 and S-5 of the Supporting Information along with the outcomes of Kolmogorov–Smirnov tests. In addition, the predicted standard deviation of the centroid distribution is plotted against the empirical one for N_p ranging from 1 to 25, in Figure S-6 of the Supporting Information. Overall the fit of the model is very good, and while the presence of a mild effect possibly arising from imperfect fitting of $f(t)$ to the voltage pulses in the R script cannot be

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

ruled out, the deviations from the predicted model are small enough that they can be attributed to statistical error.

While these results are encouraging, they do not allow us to validate all aspects of the Gaussian centroid model since the high variance of the ion arrival distribution tends to dominate smaller terms in eq 15. However, the impact of such terms can be examined with Monte Carlo simulations based on more unusual instrumental settings where otherwise negligible terms become more pronounced. In order to more unambiguously probe these terms, the Monte Carlo simulations were run with the electronic and quantization noise omitted and using the intensity invariant voltage pulse that is illustrated in Figure 4. While the Gaussian centroid model generally requires $\sigma_Z > \Delta t$ models using intensity invariant voltage pulses are more robust to departures from this assumption, and since using a low σ_Z helps accentuate the contribution to the variance of the remaining terms σ_Z was set to $\Delta t/2$. Additional settings for the simulations are listed in Table S-2 of the Supporting Information. The resulting distribution is shown on Figure 6

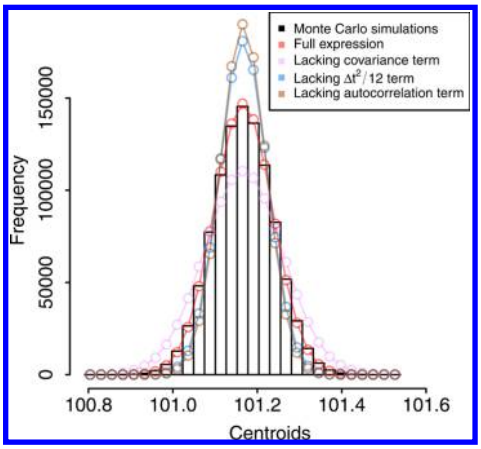


Figure 6. Comparison of the centroid distribution obtained from Monte Carlo simulations with the predicted Gaussian distribution (red) as well as three alternative Gaussian distributions obtained by removing the covariance term (pink) by removing the $\Delta t^2/12$ term (blue) or by removing the autocorrelation term (brown) from eq 15. The alternatives provide significantly poorer fits, which supports the correctness and relevance of some of the more surprising terms in the expression for the variance of the centroid distribution.

along with four Gaussian distributions, one corresponding to the predicted distribution, while the remaining three have omitted from eq 15 the covariance term, the $\Delta t^2/12$ term, and the autocorrelation term, respectively. Clearly the predicted Gaussian distribution provides by far the best fit.

Since the Gaussian centroid model is based on the ratio of two random variables, parameter values for which the denominator, the mass peak intensity, H, has an appreciable probability of producing values close to 0 can cause highly volatile behavior for which the Taylor approximations fail. In practice this is not a big constraint since the centroid should only be calculated across sampling times where a mass peak is present. Nevertheless, if the signal is weak and the electronic noise severe, the model should be employed with care. Similarly departures from Normality can be expected if the ion count is not high enough for the Central Limit Theorem to apply which is a particular risk if the signal-to-noise ratio of the signals induced by the individual ions is very low, whether this be due to their pulse height distribution or the shape of the voltage pulse. The degree of convergence to Normality that is required of the centroid distribution will depend on the manner in which this property is used and while a high signal-to-noise ratio is indicative of Normality, the adequacy of the convergence is better assessed explicitly with simulations.

In order to probe conditions under which the Gaussian centroid model breaks down, additional centroids were simulated under a range of parameter values and Kolmogorov–Smirnov tests were performed to assess whether the resulting distributions matched the predicted ones. A set of parameters corresponding to a typical electron multiplier gain[27] but with a low number of acquisitions and a weak signal were chosen as reference parameters, to reflect a presumed failure case. As such Λ was set to 1, N_p was set to 100, the shape of $f(t)$ was the same as that used for Figure 6 and illustrated in Figure 4, and μ_p and σ_p were set to 3 while ΔV, σ_s, and σ_B were all set to 1. Four subsets of these parameters were then, in turn, jointly varied across a range of values while the remaining parameters were kept fixed, along with σ_Z and Δt, which were both set to 1 throughout. The full set of parameter values used in the simulations are listed in Table S-2 of the Supporting Information. A total of 1000 centroids were simulated for each of the resulting models, and a Kolmogorov–Smirnov test was applied to each of these centroid distributions.

Figure 7 shows the resulting p-values as a function of the modified parameter values. Raising the rate of ion arrivals to a more moderate level quickly renders the Gaussian centroid model valid, as does reducing electronic and quantization noise. The model can also be rendered valid by raising the mean outcome of the pulse height distribution, and in fact convergence to it would be achieved much faster if μ_p could be increased independently of σ_p, however for most electron multipliers the two will be related. In addition, the model can be made applicable by simply increasing the number of acquisitions histogrammed, although the number required is in this case quite large due to the low rate of ion arrivals and due to the fact that histogramming more acquisitions also introduces more electronic noise. This is nevertheless a useful feature as acquisitions may be histogrammed for this purpose after the data have been written to disk. Unsurprisingly the model is never valid when there is an appreciable probability that the mass peak in question has an intensity of less than 0. Crucially, however, the range of validity is such that the model can generally be *rendered* valid by appropriate choice of settings and by appropriate engineering.

DISCUSSION

The Gaussian model of the mass peak centroid presented here appears to approximate the true centroid distribution closely under a broad range of conditions. It may fail for weak signals in the presence of high noise although this can potentially be mitigated by increasing the number of acquisitions that are histogrammed. The Gaussian centroid model makes a number of standard assumptions also made by the Gaussian intensity model, in that it requires that the mass window used does not overlap with outside mass peaks, that we are within the linear range of the electronics, and that the ADC is unsaturated at both extremes. More advanced features of the electronic noise such as electromagnetic interference or dependence across acquisitions were not considered although they could potentially be incorporated into the model if they cannot be eliminated from the experimental setup. The model is applicable to mass spectrometers that use electron multipliers

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232–2241

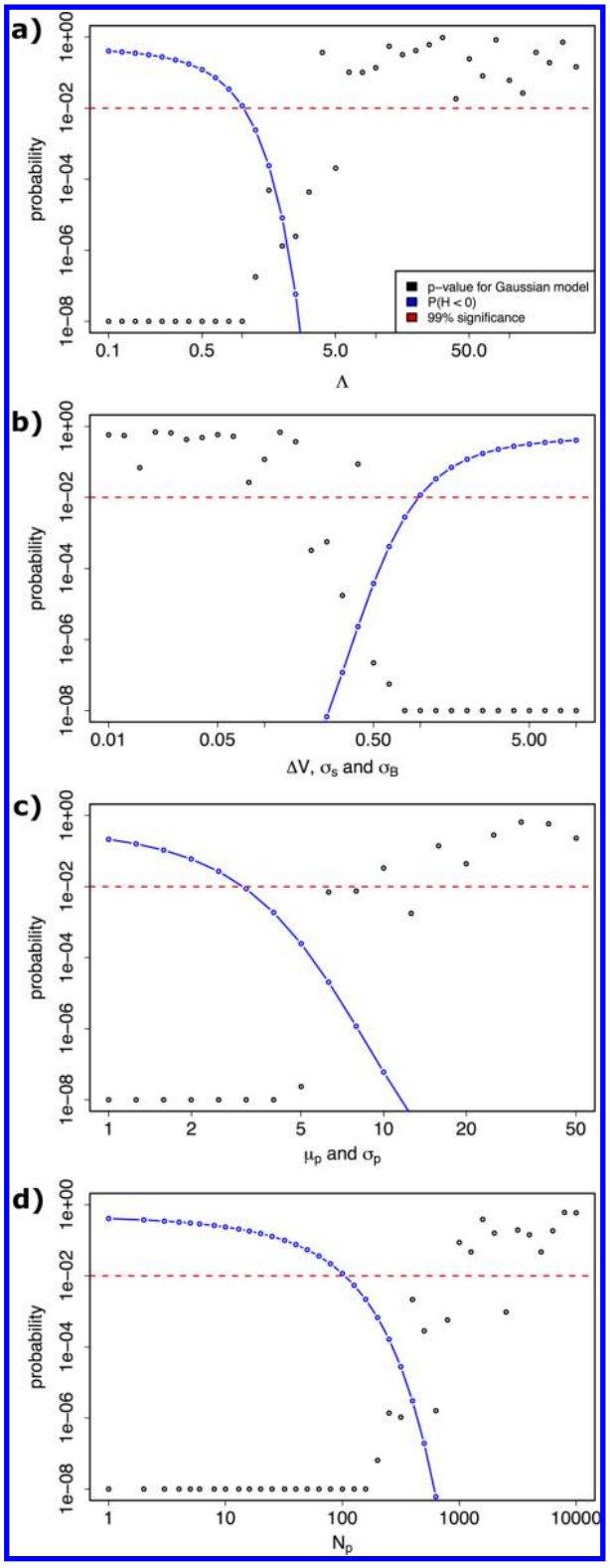


Figure 7. Evaluation of the validity of the Gaussian centroid model over a range of parameter values through the application of Kolmogorov−Smirnov tests to 1000 simulated centroids for each model. In panel a it is only the value of Λ and in panel d only the value of N_p that is altered for the different models from which the centroids are simulated. In panel b, the parameters ΔV, σ_s, and σ_B all have the same value for each model, as do μ_p and σ_p in panel c. The p-values are truncated to 10^{-8}, and the red line indicates a probability of 0.01. The blue line indicates the probability of obtaining a mass peak intensity of less than zero for a given simulation as predicted by the Gaussian intensity model. As anticipated, the Gaussian centroid model requires that this probability be negligible.

and ADCs as well as to other scientific instruments with a similar setup.

A number of extensions to the Gaussian centroid model may be required if certain of its assumptions prove difficult to satisfy. For example, if the mean baseline of the electronic noise cannot be determined with sufficient accuracy that the ADC quantization levels can be chosen so that $\mu_B = 0$, it will be necessary to write out more complicated Taylor expansions than those that have been used here. Similarly if the acquisition pulse is not well-synchronized with the ADC, μ_Z must be treated as a random variable and related to the current expressions for H and J via the law of total expectation and the law of total covariance. A more complicated extension relates to the calibration of the mass spectrometer. Since μ_Z must ultimately be related to the mass to charge ratio through a specific mapping, the parameters of that mapping must be estimated (e.g., through the use of a reference compound of known mass[25]) meaning that they too may be subject to uncertainty. The distribution of the calibrated centroid, as opposed to the raw centroid described in this article, will be impacted by any uncertainty introduced by the calibration procedure and its precise distribution (which may be non-Gaussian) must be worked out from the particular form of calibration procedure used. A strong case can be made for choosing this procedure so that the calibrated centroid has a simple and tractable distribution, whose bias and variance can be determined and minimized. If this proves possible it will likely also be possible to construct a statistical test of hypothesis to determine whether an observed centroid is consistent with the theoretical m/z of a putative molecular formula. This would not only constitute a major milestone in MS data analysis but could also provide a powerful guide for the construction of future mass spectrometers since the impact that design choices would have on the statistical power of such a test could serve as a measure of their analytical utility.

While innovations such as orthogonal acceleration and the use of reflectrons have enabled high mass accuracy for TOF instruments, further improvements in mass accuracy on the basis of ion optics alone are likely to be more modest. However, the $(N_p\Lambda)^{-1}$ dependence in the expression for the variance of the centroid suggests that further reductions in the variance of estimates of μ_Z can be achieved simply by collecting more ions (e.g., through repeated experiments), while avoiding saturating the ADC. The idea of pooling multiple measurements is not new. However, in order for it to enable significant improvements in mass accuracy it must be combined with a reduction in the bias of the individual estimates,[28] or else the pooled estimate will converge to the wrong value. Therefore, an alternative avenue to attaining high mass accuracy might be sought through the very careful consideration of biases, such as that due to the pulse-peak correlation, which once identified may in principle be accounted for. Such an approach would ultimately have to address any biases inherent to the ion optics as well as any introduced as part of the calibration. These are difficult tasks that will likely require more extensive efforts at mathematical modeling of mass spectrometry fundamentals; however, the central reliance of this approach on high ion

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241

counts would fit well with the high sensitivity of TOF and quadrupole instruments.

CONCLUSION

All of the above are ambitious extensions that could do much to strengthen the statistical rigor of MS data analysis and to extend the scope of the inferences that can be drawn from it. Whether they are feasible or not depends on the feasibility of engineering analytically tractable mass spectrometers that produces data whose probability distribution can be described mathematically. Resolving this question is unfortunately difficult due to the varied subcomponents that go into a mass spectrometer and the associated division of research. There are few research groups that will have both thorough expertise in, and influence over the design of, ion optics, electron multipliers, preamplifers, and ADCs all of which might have to be investigated and fine-tuned in an integrated manner in order to probe the limits of this approach. This division of research may be part of the reason why the rather poorly understood statistical fundamentals of mass spectrometry have not hitherto prompted much research on the topic, despite the acknowledged importance of statistical rigor across the sciences. A possible first step toward addressing this obstacle would be for MS vendors to provide more extensive software support for the acquisition of raw data and for the determination and manipulation of fundamental instrumental parameters such as those relating to the pulse height distribution. While this would not directly resolve the engineering challenges it could do much to facilitate research that seeks to lay the needed theoretical groundwork and it would begin to make the field accessible to the many laboratories lacking the specialized equipment currently needed.

ASSOCIATED CONTENT

Supporting Information

Supporting Information available as noted in the text. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02446.

> Elementary properties under idealized sampling; multiple ion arrivals with variable pulse heights; electronic noise and quantization errors; histogramming; mean and variance of C; intensity invariant voltage pulses; model parameters; $g(t)$ and $f(t)$ curves; ^{36}Ar spectra; Gaussian distributions, quantile−quantile plots; and the predicted and the empirical standard deviation for the centroid distributions (PDF)

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Andreas Ipsen: 0000-0002-2566-8811

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

The author wishes to thank Gareth Brenton and Richard Smith. This work was supported by the MRC through Grant No. MR/J013994/1.

REFERENCES

(1) 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, 2001.
(2) Opsal, R. B.; Owens, K. G.; Reilly, J. P. *Anal. Chem.* **1985**, *57*, 1884−1889.
(3) Guilhaus, M. *J. Mass Spectrom.* **1995**, *30*, 1519−1532.
(4) Coles, J. N.; Guilhaus, M. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 772−778.
(5) Coombes, K. R.; Koomen, J. M.; Baggerly, K. A.; Morris, J. S.; Kobayashi, R. *Cancer Inform.* **2005**, *1*, 41−52.
(6) Peregudov, O. N.; Buhay, O. M. *Int. J. Mass Spectrom.* **2010**, *295*, 1−6.
(7) 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.
(8) Peterson, D. W.; Hayes, J. M. *Contemporary Topics in Analytical and Clinical Chemistry*; Hercules, D. M., Hieftje, G. M., Snyder, L. R., Evenson, M. A., Eds.; Plenum Press: New York, 1978; pp 217−252.
(9) Coates, P. B. *J. Phys. E: Sci. Instrum.* **1968**, *1*, 878−879.
(10) Coates, P. B. *Rev. Sci. Instrum.* **1992**, *63*, 2084−2088.
(11) Stephan, T.; Zehnpfenning, J.; Benninghoven, A. *J. Vac. Sci. Technol., A* **1994**, *12*, 405−410.
(12) Ipsen, A.; Ebbels, T. M. D. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 779−791.
(13) Ipsen, A.; Ebbels, T. M. D. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1824−1827.
(14) Lombard, F. J.; Martin, F. *Rev. Sci. Instrum.* **1961**, *32*, 200−201.
(15) Ipsen, A. *Anal. Chem.* **2015**, *87*, 1726−1734.
(16) Nikolaev, E. N.; Kostyukevich, Y. I.; Vladimirov, G. N. *Mass Spectrom. Rev.* **2016**, *35*, 219−258.
(17) Makarov, A. *Anal. Chem.* **2000**, *72*, 1156−1162.
(18) Dahl, D. A. *Int. J. Mass Spectrom.* **2000**, *200*, 3−25.
(19) Garimella, S. V. B.; Ibrahim, Y. M.; Webb, I. K.; Tolmachev, A. V.; Zhang, X.; Prost, S. A.; Anderson, G. A.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1890−1896.
(20) Casella, G.; Berger, R. L. *Statistical Inference*; Duxbury: Pacific Grove, CA, 2002.
(21) Vaart, A. W. van der. *Asymptotic Statistics*; Cambridge University Press: New York, NY, 2000.
(22) Wannamaker, R. A.; Lipshitz, S.; Vanderkooy, J.; Wright, J. N. *IEEE Trans. Signal Process.* **2000**, *48*, 499−516.
(23) Guilhaus, M.; Selby, D.; Mlynski, V. *Mass Spectrom. Rev.* **2000**, *19*, 65−107.
(24) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849−865.
(25) Wolff, J.-C.; Eckers, C.; Sage, A. B.; Giles, K.; Bateman, R. *Anal. Chem.* **2001**, *73*, 2605−2612.
(26) Knoll, G. F. *Radiation Detection and Measurement*; John Wiley & Sons: New York, 2010.
(27) Rather, O. *Saturation correction for ion signals in time-of-flight mass spectrometers*. U.S. Patent Application 20130181123 A1, July 18, 2013.
(28) Gedcke, D. A. *Suppressing Noise in TOF-MS with FASTFLIGHT-2*; AN62, Ortec: Oak Ridge, TN, 2001.

DOI: 10.1021/acs.analchem.6b02446
Anal. Chem. 2017, 89, 2232−2241