

$S_{(X)} = (s_{(X)i})_m$	Euclidean distances of test-set replicates from C		bootstrap distribution; CDF has $(2m - 4pm)$ elements
$S_{(V)} = (s_{(V)i})_m$	Euclidean distances of validation-set replicates from C	$C_{(X)} = (c_{(X)i})_{2m-4pm}$	CDF formed by the trimmed and ordered elements of the test-set and training-set bootstrap distributions
$P_{(T)} = (p_i)_{m-2pm}$	set of $(m - 2pm)$ indices used for trimming distance distributions	$C_{(V)} = (c_{(V)i})_{2m-4pm}$	CDF formed by the trimmed and ordered elements of the validation-set and training-set bootstrap distributions
$C_{(t)} = (c_{(t)i})_{2m-4pm}$	cumulative distribution function (CDF) formed by the trimmed and ordered elements of the training-set		

Quantile Analysis: A Method for Characterizing Data Distributions

ROBERT A. LODDER* and GARY M. HIEFTJE†

Department of Chemistry, Indiana University, Bloomington, Indiana 47405-4001

Analyzing distributions of data represents a common problem in chemistry. **Quantile-quantile (QQ)** plots provide a useful way to attack this problem. These graphs are often used in the form of the normal probability plot, to determine whether the residuals from a fitting process are randomly distributed and therefore whether an **assumed** model fits the data at hand. By comparing the integrals of two probability density functions in a single plot, QQ plotting methods are able to capture the location, scale, and skew of a data set. This procedure provides more information to the analyst than do classical statistical methods that rely on a single test statistic for distribution comparisons.

Index Headings: Near-infrared, Chemometrics.

INTRODUCTION

The field of analytical chemistry abounds with examples of data-distribution analysis. Analytical chemists have utilized or studied it in such diverse areas as the two-dimensional distribution of electrochemical activity on graphite-epoxy electrodes,¹ the sequence distribution of ethylene-propylene copolymers using carbon-13 NMR and Markovian **statistics**,² the distribution of crystallographic structure factors in the course of making accurate measurements of crystal **structures**,³⁻⁵ and even the **distribution** of components in empirical mathematical models used in the study of the homogeneity of **solids**.⁶ In large part, these distributions are derived from a **proposed** model for a chemical or physical phenomenon; many times, the purpose of an experiment is to compare an expected theoretical distribution of data to some distribution of observations actually obtained in the **laboratory**. In the end, the distributions are either shown to be identical, thus verifying the theoretical model of the phenomenon, or they are shown to be different in some way that suggests that a new and different model more appropriately describes the data.

Often, the initial comparison of univariate populations

involves merely a visual inspection of the frequency distributions of each data set, usually through the use of histograms,⁷ frequency **polygons**,⁸ or **stem-and-leaf**⁹ displays. Visual inspection is a very important part of the process of distribution analysis because human beings possess a remarkable ability to grasp visual patterns and trends, even with very little prior knowledge of what to expect in the data. Simple visual inspection can suggest, among the data, relationships that were previously unexpected. Often, numerical restatements of the data actually tend to have the opposite effect—that is, numerical methods frequently conceal important structures in data when the methods were not intentionally designed to detect these structures.

The cumulative frequency distribution is often used in constructing these graphical displays because of the ease with which it indicates the number of values above or below a selected point or observation or, conversely, the point which marks the beginning or end of a chosen fraction of all random variables. This curve, which is just a plot of the integral of the probability density function, is unfortunately often too unwieldy for a simple visual analysis, and for this reason (as well as for others discussed later) **its** use is eschewed by many workers. What would be most useful is a technique that retains the convenience of the cumulative distribution function (CDF) in locating the boundaries of specified areas, but that also transforms the data from a complex curve with inflections into a more familiar linear form. This feature, in fact, is the essence of the quantile-quantile plot.

The following review of distribution analysis outlines many procedures that have been used in chemistry. The use of analysis by distribution quantiles is also developed from basic principles of statistics. Finally, quantile analysis is shown to be more flexible than typical numerical test procedures in representative applications to chemical analysis.

THE NATURE OF QUANTILE ANALYSIS

Quantile-quantile (**QQ**) plots are used to make detailed comparisons of two collections of data. They are typically

Received 3 June 1988.

* Present address: College of Pharmacy, University of Kentucky, Lexington, KY 405364082.

† Author to whom correspondence should be sent.

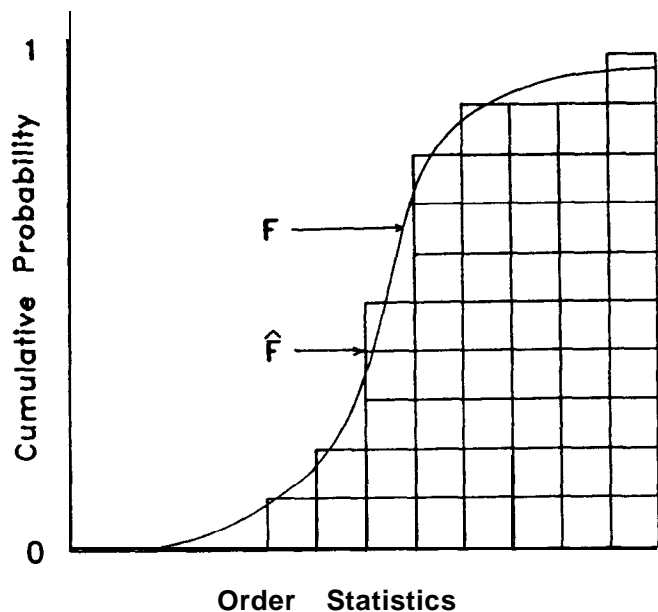


FIG. 1. Constructing an ECDF. The blocks represent observations. As the number of observations increases, the rough empirical function, \hat{F} , approaches the smooth TCDF, F . Order statistics are the rank-ordered observations, from lowest to highest.

created by plotting the empirical cumulative distribution function (ECDF), $F(\mathbf{1})$, of one set of data against the ECDF, $F(Z)$, of the other. The ECDF can be viewed graphically as n independent random variables, or observations, represented as building blocks and stacked along a horizontal axis so that their running sum forms an ever-increasing "staircase" (see Fig. 1). The steps in the staircase have a net height of $1/n$ over each observation, and the total height of the stairs ultimately reaches a value of one (i.e., n/n).¹⁰ The intersection of a horizontal line through a given cumulative probability with a vertical line through the ordered data gives a plotting position, or a quantile, for the QQ plot. This use of the ECDF in a QQ plot does not necessarily depend upon any assumption of a particular parametric distributional specification (for instance, this sort of test need not assume that the empirical distribution is Gaussian), and allows the QQ plot to be a powerful and flexible tool in exploratory data analysis.

Quantile-quantile plots are commonly used to verify the distributional properties of a set by comparing the set of observations against a pre-specified "model" distribution," and to obtain insight into the nature of the "true" distribution underlying the experimental observations with respect to some reference distribution (i.e., the actual observations have asymmetry, heavier tails, lighter tails, etc., when compared with some standard distribution). In a particular sense, this power is often exploited in the analysis of residuals from a modeling process.

For example, if one wished to determine whether the Beer-Lambert law held for near-infrared diffuse reflectance spectrometry, one might perform a linear fit on near-IR calibration-sample spectral data. A linear fitting process like ordinary least-squares (OLS) regression produces some sort of straight line regardless of how well the data actually conform to the linear assumption, so

some sort of analysis of the goodness-of-fit must be undertaken. The correlation coefficient can be calculated for the fit, of course, and this coefficient provides a useful indication of how well the line actually describes the observations. With the use of a single-valued statistic like the correlation coefficient as a goodness-of-fit criterion, parameters of a prediction model can be easily adjusted and the correlation coefficient recalculated in order to obtain the best parameters for prediction of future values. However, because the classical linear model is given by

$$y(i) = m(1)x(i, 1) + \dots + m(j)x(i, j) + e(i) \quad (1) \\ (i = 1, \dots, n),$$

where the error $e(i)$ is usually assumed to follow the Gaussian distribution with zero mean and unit variance, a more general fitting method can profitably be employed. In this method the parameters are adjusted until the $e(i)$'s become randomly distributed. This sort of modeling procedure has become commonplace in the statistical literature; in the case of the ordinary least-squares example, the procedure would be applied to the set of near-IR observations with the use of the linear model created with the OLS technique to form a new set of observations, the residuals. By definition of the general fitting method, a QQ plot of these residuals (on the ordinate) vs. the quantiles of the Gaussian distribution (on the abscissa) would then form a straight line of unit slope through the origin if the linear model holds.

Many other nonparametric methods used to analyze residuals are concerned only with the sign of the residual value, and treat it as though only two values were possible. In contrast, quantile analysis uses both the magnitude and sign of the residuals to determine whether the residuals match a selected distribution; yet the method retains its overall nonparametric character. This ability is of particular importance when robust or median-based fits (alternative fitting procedures not based on a Gaussian minimization of the sum of the squares of the residuals) are being employed. These robust procedures are frequently employed when certain assumptions, required for ordinary least-squares fitting to be valid, are not met. For example, one such assumption is that the data points are normally distributed with constant variance. If the OLS algorithm and one of these robust fitting procedures are both applied to the same set of bivariate data, the residuals from the model produced by the OLS algorithms will tend to appear to be more normally distributed than the residuals from the robust method. The relative absence of this "masking effect" in robust algorithms makes quantile analysis of their residuals even more important because of the increased probability of extracting useful information about the structure of the data from the distribution of the residuals. The power of the QQ plotting procedure becomes especially evident when one realizes that it is just as applicable to higher-order, nonlinear modeling problems.

Physicists have exploited the modeling capability of QQ plots in the analysis of positron-annihilation angular correlation and Compton-profile experiments.¹² These experiments are used to study the electron momentum distribution in materials, and the changes in the distribution profiles often are quite small and near the level

of statistical uncertainty. In the past, a one-parameter chi-square test has been used in the comparison of data profiles, but this technique requires a good estimate of the standard deviations of the distributions before it produces meaningful results. Unfortunately, making a good estimate of the standard deviation is sometimes impossible—as when multiple scattering effects are present in Compton-profile experiments. QQ plotting procedures do not require a priori estimates of scale parameters, and can be used to suggest changes in an inadequate model rather than merely to reject it.

QUANTILE THEORY

In order to understand better the power, flexibility, and utility of quantile plotting, we must more closely examine its theoretical development. Given two distribution functions, $F(1)$ and $F(2)$ (see Fig. 2), a quantile plot can be formed by creating data pairs from abscissa (X_i) values that are the p th quantile of $F(1)$, and ordinate (Y_i) values that are also the p th quantile of $F(2)$ [where p is the cumulative probability—one data pair is formed for each value of p on the interval $(0, 1)$]. Put simply, the QQ plot of $F(1)$ against $F(2)$ is a plot of the X_i and Y_i points corresponding to a particular limit of integration as that cumulative probability p is allowed to vary between zero and one.

One could use a plot such as Fig. 2 to compare an ECDF [For $F(1)$] to an actual CDF [F or $F(2)$]. Naturally, as n increases, the ECDF converges to the actual CDF of the process responsible for generating the observations. In fact, plots such as Fig. 2 have been used in actual distribution comparisons of particle sizes obtained by turbidimetric measurements¹³ and in the estimation of distribution functions and determination limits of ultraviolet-absorbing species in plant extracts.¹⁴ In the case of the turbidimetric measurements, light scattering by a suspension of fine particles was used to determine the mean and range of the particle diameters. It was assumed that this empirical distribution could be fitted to an ordinary two-parameter log-normal distribution function, and the parameters were estimated from the light-scattering data obtained at two different wavelengths. The cumulative distribution plots gave results that were “reasonably representative” of the central portion of the distributions.

The study of the distribution functions of ultraviolet-absorbing substances in plant extracts¹⁴ was performed with the use of high-performance liquid chromatography. The chromatograms of 62 plant-leaf extracts were used as empirical distribution functions. A computer simulation was also performed to estimate the true distribution functions of the absorbances of observed peaks. These distributions were then used to assess the probability that a given plant constituent could be successfully determined.

In general, CDF plots such as Fig. 2 test the location, shape, and scale of the empirical distribution. To use such a plot to decide whether or not a given set of variables follows a specified CDF, the location and scale parameters of the distribution must first be estimated. This estimation process can range from being a difficult process to a nearly impossible one. Such was the situation

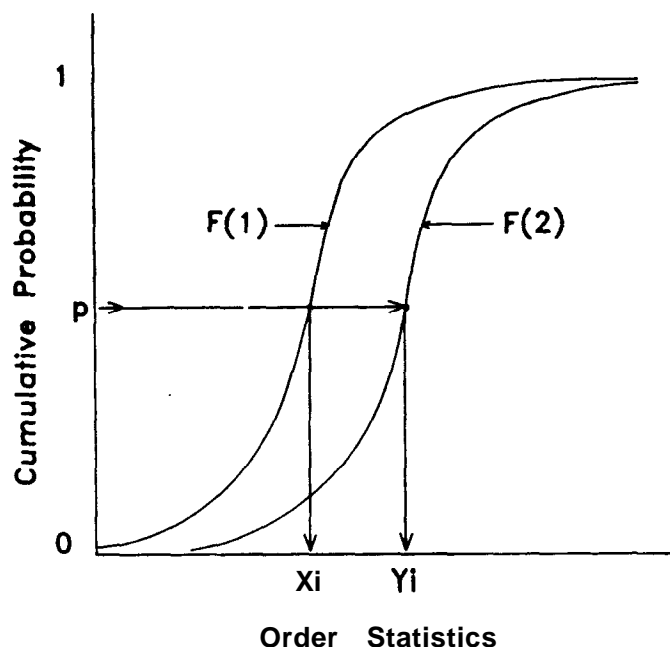


FIG. 2. Forming quantiles for plotting from CDFs. Formulas determine a series of p values to use as plotting positions, or quantiles $\{q(p)$ or $(X_i, Y_i)\}$. Each p gives a pair of order statistics to use as a point in a QQ plot.

in the case cited earlier for Compton-profile experiments.* In many cases, neither the chi-square method nor the simple CDF method provides a satisfactory solution to the distribution-analysis problem.

Another problem with simple CDF plots is that they often fail to reveal significant variations in curves near $p = 0$ and $p = 1$ -regions which are often of critical concern. For instance, when one investigates whether an ECDF is normal, the outliers are of major importance. If outliers contaminate 10% of the data, then the regions between zero and the fifth percentile and between the ninety-fifth and one hundredth percentiles are of great interest. A plot that decompresses the data in these regions possesses the greatest utility.

QQ plots do not suffer from these data-compression and parameter-estimation problems. By obtaining quantiles from CDF plots and plotting them vs. each other [i.e., $F^{-1}(1)(p)$ on the y-axis and $F^{-1}(2)(p)$ on the x-axis], one can avoid these difficulties. The notation $F^{-1}(p)$ simply indicates the empirical inverse of the cumulative distribution function given by the order statistics.

The probability plotting positions (p) are not chosen randomly to give quantiles for plotting.¹⁵ Equations have been developed to give the best positions for families of cumulative distribution functions. The fact that ECDFs are discrete makes them necessarily somewhat ambiguous. As a result, sampled ECDF points for simplicity are composed of elements of the set of order statistics X (the rank-ordered experimental data): The corresponding points of the theoretical CDF (TCDF) are $F^{-1}(p)$ where $p = (i - 0.5)/n$ (see Ref. 15) or some other cumulative probability position [such as $p = (i - 0.4)/(n + 0.2)$; see Ref. 16].

This quantile-selection procedure yields a straight line with unit slope through the origin if $F(1)$ is identical to $F(2)$ (or if F is identical to F). The reason for this linearity

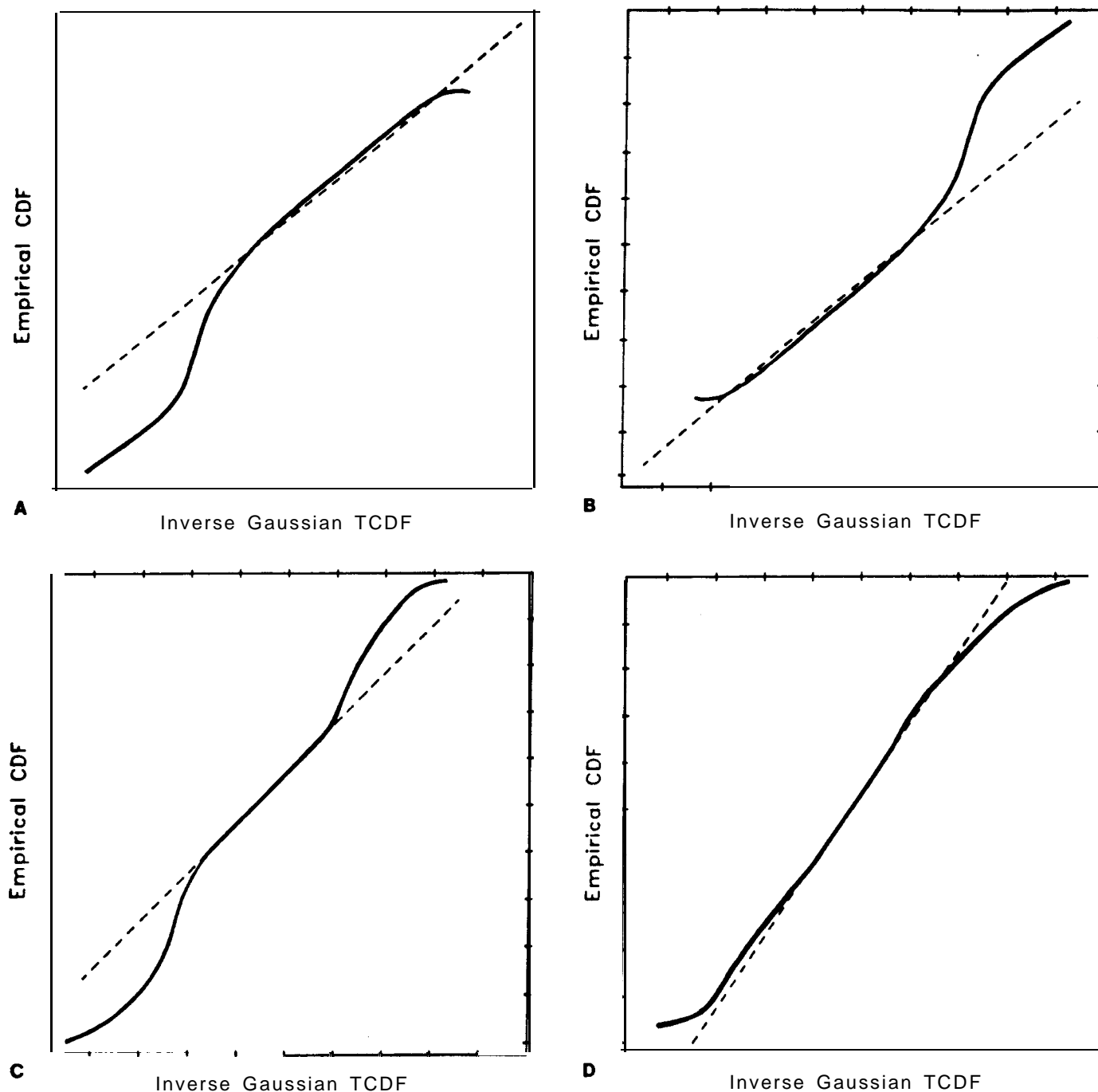
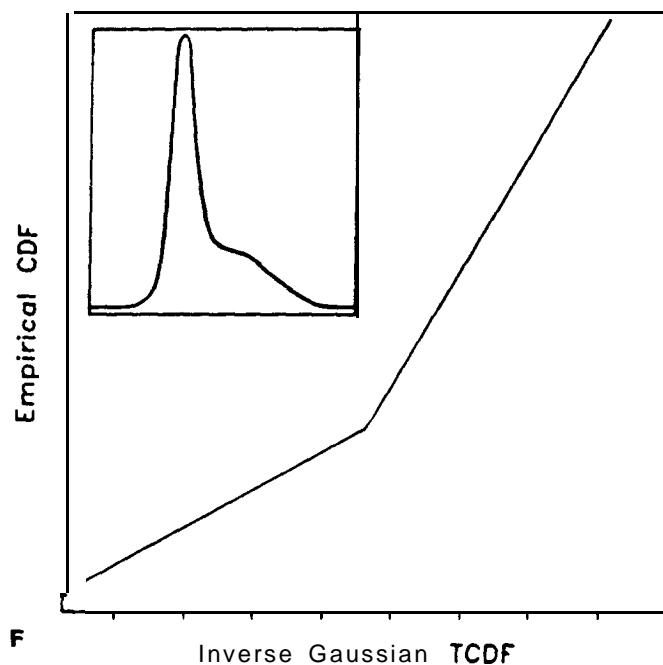
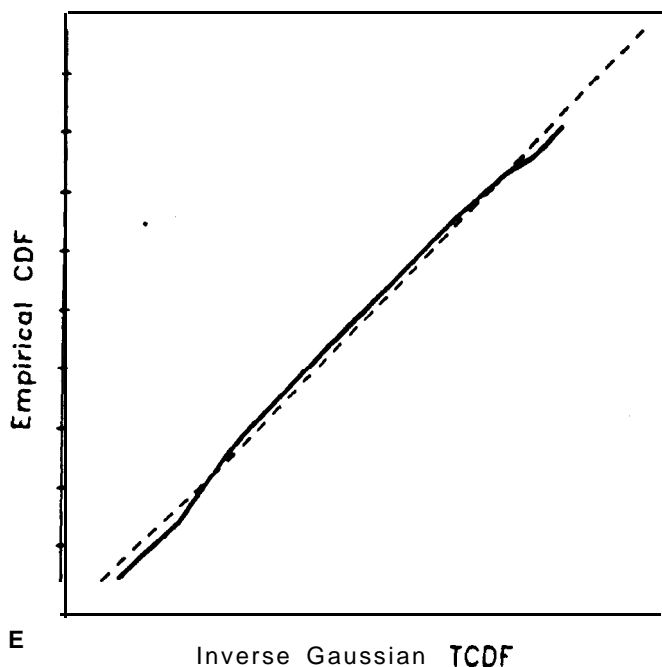


FIG. 3. Normal probability plots of synthetic inverse Gaussian spline functions vs. the standard inverse Gaussian function, $\Phi^{-1}(x)$. The standard Gaussian is on the abscissa. The plots have probability axes. The best straight line through each has an intercept equal to the mean and a slope equal to the standard deviation of the ECDF on the ordinate. (A) Right-skewed. The pattern shows a distribution on the ordinate that is initially "heavier" in observations than the reference standard Gaussian. (B) Left-skewed. This situation can be thought of as the reverse of 3A. Left skew indicates that the lower observations are exhibiting a leverage effect on the mean of the data distribution. (C) Heavy-tailed. This situation can be viewed as being both right-skewed and left-skewed. (D) Light-tailed. This situation can be thought of as a truncation of the right end left tails of the data distribution. This case is the inverse of the heavy-tailed case. (E) Gaussian with different location and scale. The ECDF Gaussian has a mean and variance of 4. The best straight line through these points therefore has a slope of 2 and an intercept of 4. The relative straightness of the line indicates that no skew is present in the empirical data. (F) Bimodal (two unresolved Gaussians). Two distinct lines can be fit to these data. The slopes and intercepts of the two lines define the two Gaussian curves shown in the inset.

is simple: If two functions $[f(1)$ and $f(2)]$ of x are identical, a plot of their integrals $[F(1)$ and $F(2)]$ from negative-infinity to x , as x is allowed to vary through the domains of the functions, will produce two identical vectors of points $[F(1),i$ and $F(2),i]$. Plotting $F(1),i$ vs. $F(2),i$

will give a straight line. If $F(1)$ differs from $F(2)$ by only a location and/or scale change, the plot will still be a straight line, but with a slope and intercept that depend upon the values of the location (μ) and scale (σ) changes. [Note that location and scale changes are normalized for



comparisons between CDFs by a simple relationship: If $F(2)(x) = F(1)[(x - \mu)/\sigma]$, the slope of the line formed will be σ and the intercept will be μ/σ .)

If $F(1)$ differs from $F(2)$ in a more fundamental way, the QQ plot will no longer be a straight line. Figure 3 shows some of the possible patterns that can emerge. The plots given in Fig. 3(A-F) assume that the functions $F(1)$ and $F(2)$ are given by smooth curves, but this assumption is merely for simplicity. The curves were generated by creating distribution functions that show certain properties with respect to the Gaussian distributions: right-skewed, left-skewed, light-tailed, heavy-tailed, Gaussian but with different location and scale parameters, and bimodal (the sum of two Gaussians). The random variables selected for plotting were obtained through Monte Carlo integration of the probability density functions of both the created and Gaussian distributions. The definition of a QQ plot is equally valid when $F(1)$, $F(2)$, or both are step functions, like histograms.

APPLICATION OF QUANTILE ANALYSIS

The interpretation of patterns such as those in Fig. 3 has been of analytical interest in the field of biochemistry. A bimodal distribution composed of two Gaussians gives a QQ plot to which two distinct straight lines can be fitted (see Fig. 3F)—this fact has been used to assign “normal-healthy” blood-glucose values in human test populations, and to identify defective control of blood sugar.¹⁷ Similarly, deoxyribonucleic acid (DNA) melting-point data have been plotted to allow the mean content of guanine plus cytosine in DNA to be determined, and to provide insight into the degree of variation from the mean in DNA-base composition of fifteen different bacterial strains.¹⁸ In this technique, a DNA sample is slowly heated until a sharp increase in absorptivity is observed. The temperature at which this transformation occurs signals the double-stranded DNA helix breaking apart

to form a single-stranded random coil. The midpoint of the rise in absorptivity is called the thermal melting value, or T_m , and is directly proportional to the sum of the guanine and cytosine present in the DNA sample.

The ordinary method of making T_m determinations involves plotting absorptivity vs. temperature, and results in a plot resembling a CDF plot, or simply the titration curve of a weak acid by a strong base. The midpoint of the rise, T_m , can be found in a manner similar to that employed in locating the equivalence point in a routine titration-curve analysis.

In contrast, the QQ plotting method for determining T_m values is based on the assumption that the compositional distribution of nitrogenous bases in DNA is Gaussian. The “CDF” plot of absorptivity vs. temperature can be combined with the TCDF of the Gaussian distribution to produce a QQ plot. If the distribution of bases is indeed Gaussian, this fact will lead to the formation of a straight line on the QQ plot, and the T_m value will be obtainable from the fiftieth percentile point [where $\Phi(0) = 0.5$]. Not only is this procedure visually easier than estimating an equivalence point, but it also requires fewer data points (as few as four, according to Knittel¹⁸), permits a more rapid determination of the standard deviation of T_m , and, by virtue of the fact that QQ plots give a straight line for each component distribution, makes it possible for the melting procedure to reveal the presence of minor components and impurities as well.

QQ plots can also be easily generated directly from histograms. Figure 4 is the result of two multistep functions, one a theoretical CDF approximating the Gaussian distribution and placed on the x-axis, and the other, on the y-axis, an ECDF generated by a set of independent observations following an “unknown” distribution law that is asymmetric and was created to have a mean of 14.6 and a standard deviation of 5.5. This sort of plot is known as a normal probability plot, and it gives a slope equal to the standard deviation of the ECDF and an

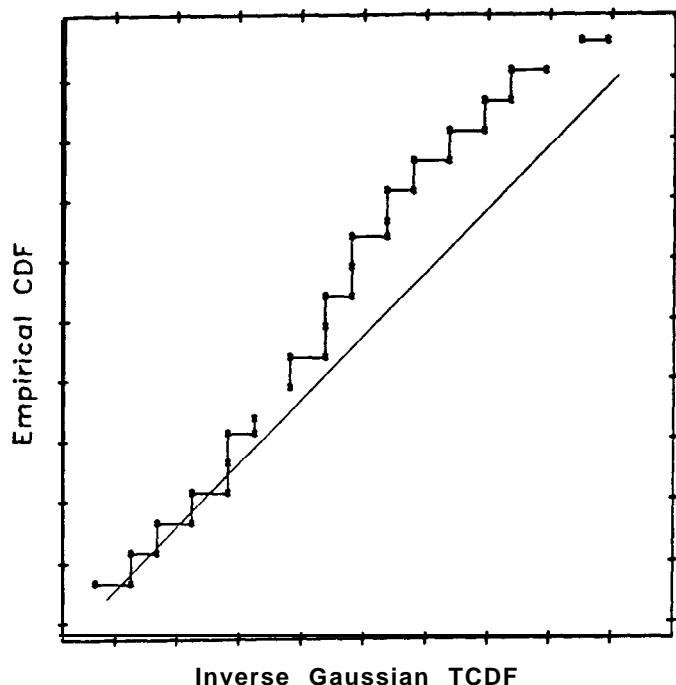


FIG. 4. Normal probability plot of histograms. The lack of smoothness, or the "stair-like" quality to the plot, is due to the use of histograms for the standard Gaussian on the abscissa and the ECDF on the ordinate.

intercept equal to the mean. Normal probability plots have, to date, been the principal application of QQ plotting procedures in analytical chemistry. Their use has been most extensive in crystallography, as a result of their early introduction to that field by Abrahams and Keve.¹⁹

In Fig. 4 the quantiles $\{q(p)\}$ of the ECDF formed by the set of independent observations $x(1), x(2), \dots, x(n)$ are plotted on the ordinate of the QQ plot, and are given by

$$q(p) = x(i) \quad \text{for } [(i-1)/n] < p \leq i/n, \quad (2)$$

$$(0 < i \leq n-1)$$

and

$$q(p) = x(n) \quad \text{for } p = 1 \quad (3)$$

where $x(1), x(2), \dots, x(n)$ represents the order statistics for the n observations. The quantiles for $p = 1/n, 2/n, \dots, 1$ are plotted because the quantiles of the ECDF do not change as the cumulative probability varies from $(i-1)/n$ to i/n . The stair-like quality of Fig. 4 arises from the discontinuities introduced by the histograms, and it vanishes as the number of increments, n , approaches infinity.

In practice, $p = (i-0.5)/n$ is often used to give $q(p)$ (the quantile for plotting) for the following reason: If the data are a random sample from the distribution function F , the value of $x(i)$ will by definition tend to line up on $y = x$ when the ECDF given by $E(F)[x(i)]$ is plotted. (E represents the expectation value of F .) This is because the plots are based on the approximate relationship

$$E(F)[x(i)] \approx F^{-1}[(i-0.5)/n]. \quad (4)$$

A simple calculation shows why such a basis is valid. Suppose $n = 17$ and $i = 9$. The median $\{$ given by the

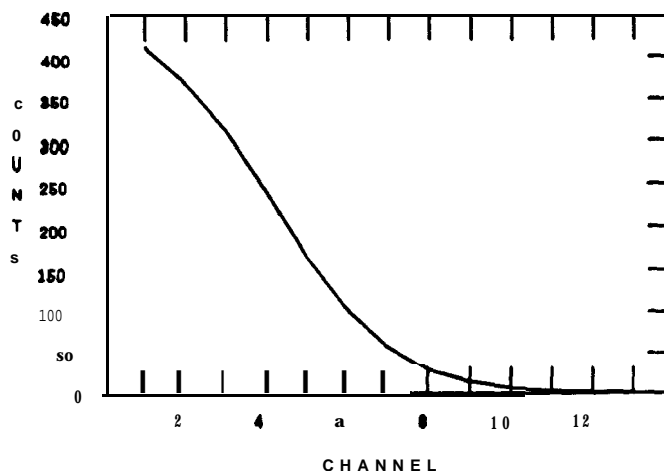


FIG. 5. The upper half of a hypothetical Thomson-scattering profile. Signal intensity on the ordinate is given in terms of photon counts, and the wavelength shift on the abscissa is scaled to detector array elements. The shape of the distribution is Gaussian.

order statistic $X\{(n+1)/2\}$ of the sample is then $x(9)$, so it is expected that one-half of the observations will be below $x(9)$. This corresponds to $p = 1/2$ because

$$(i-0.5)/n = 8.5/17 = 1/2. \quad (5)$$

Other plotting approaches, such as the more robust $p = (i-0.4)/(n+0.2)$, give similar results on analysis:

$$p = (9-0.4)/(17+0.2) = 8.6/17.2 = 0.5. \quad (6)$$

In contrast, $(i-1)/n$ gives $p = 0.47$. The more robust plotting positions do not change the appearance of the plot appreciably, but they do allow robust goodness-of-fit test statistics to be calculated directly from the plots.²⁰

The straightforward relationship between p and the order statistics suggests a simple way of analyzing QQ plots that appear similar to that given in Fig. 4, for example. In this method, one selects a quantile, for example the fiftieth percentile of the Gaussian distribution on the abscissa, which corresponds to $x(i) = 0$. The corresponding point on the plot does not lie on the line $y = x$, but rather lies above it. This particular p selection is further through the ECDF on the ordinate than the same quantile of the TCDF on the abscissa, and indicates not only that the ECDF is asymmetric but also that it is "heavier" in larger values than is the reference distribution.

Thomson scattering in the inductively coupled plasma (ICP) provides one example that clearly demonstrates the ability of quantile analysis to reveal important details about the shapes of data distributions. When the plasma is irradiated by an external source, Thomson scatter is seen as the Doppler shifting of radiation incident on the plasma by species present in the plasma. Electrons exhibit the dominant Doppler shifts because their low mass enables them to be accelerated to the highest velocities; hence, Doppler shifts can be used to determine the electron temperature of an ICP. Figure 5 shows the upper half of a hypothetical Thomson-scattering experiment using an array of detectors to monitor the Doppler shifts of a laser pulse centered on channel 0. To describe it simply, the faster the electrons in the plasma move, the higher the electron temperature of the plasma is and the broader the bell-shaped distribution

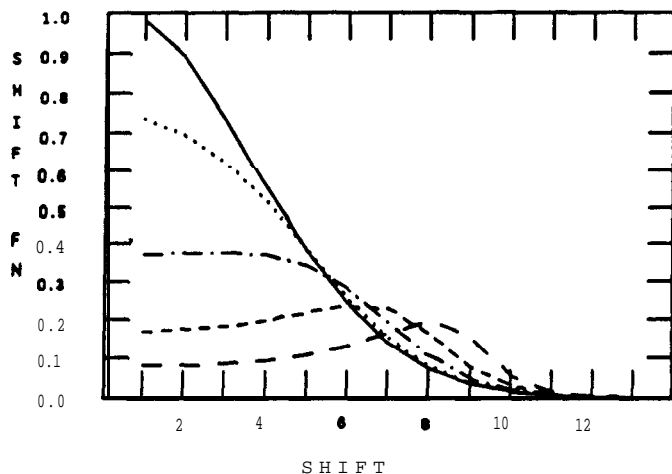


FIG. 6. Theoretical Thomson-scattering profiles corresponding to $\alpha = 0$ (solid line), $\alpha = 0.4$ (dotted line), $\alpha = 0.8$ (dot-dashed line), $\alpha = 1.2$ (short-dashed line), and $\alpha = 1.6$ (long-dashed line). The shift function is Γ from Ref. 22.

of shifted radiation becomes. Unfortunately, other factors **also** influence the shape of the shifted-radiation distribution. One such factor is the scattering parameter α .^{21,22} When $\alpha = 0$, the distribution of the shifted radiation is Gaussian (see Fig. 6).

The simplest method of determining the electron temperature from the shifted radiation in the ICP "linearizes" the Gaussian by taking the logarithm of both sides of the Gaussian function. A plot of $\ln(\text{signal})$ vs. the square of the shift then gives a line whose **slope** can be used to calculate the electron temperature. However, when $\alpha > 0$, the "linearized" data show a nonlinear trend (see Fig. 7) that causes a systematic error in electron temperatures determined by this method.

Quantile analysis allows the shape of a curve to be determined simultaneously with its scale. By using the I' function in Ref. 22 as the TCDF and adjusting a until the line in the QQ plot is straight (see Fig. 8), one can determine the shape of the experimental curve by matching it to the TCDF (the TCDF would be a member of the family of curves shown in Fig. 6). The scale ("standard deviation") can then be estimated (relative to the TCDF) from the slope of the line in the QQ plot.

Quantile plotting has been applied in our own laboratory to near-infrared reflectance spectrometry. **Near-IR** spectrometry is a rapid analytical technique that typically uses the diffuse reflectance of a sample at several wavelengths to determine the sample's **composition**.²³ Through a computerized modeling process (generally employing multiple linear regression), near-IR **spectrometry** is able to correct automatically for background and sample-matrix interference, making ordinarily difficult analyses seem routine. This modeling process employs a "training set" of samples to, in effect, "teach" the computer to recognize relationships between minute spectral features and sample composition. Of course, the training set must have been previously analyzed by some other reliable (reference) chemical procedure.

The near-IR calibration model that is developed is composed of linear equations that relate sample composition to the weighted sum of the **reflectances** at the monitored wavelengths. Unfortunately, any amount of

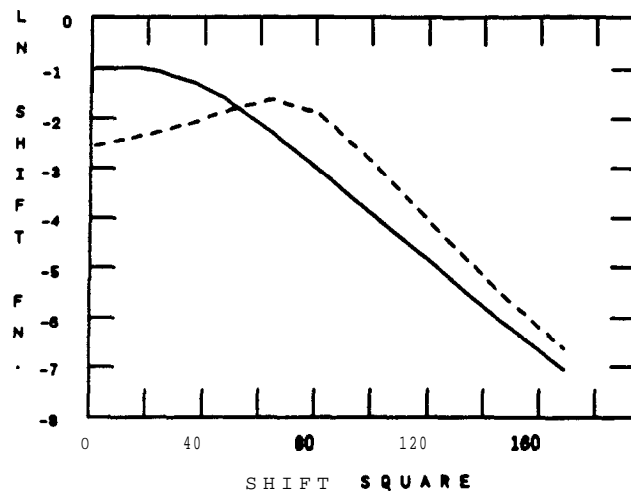


FIG. 7. The effect of attempting to linearize the Thomson-scattering profile when $\alpha = 0.8$ (solid line) and $\alpha = 1.6$ (dashed line). Calculations based on linear fits to these lines produce biased electron temperatures.

reflectance at the selected analytical wavelengths generates a corresponding composition value, regardless of the material responsible for the reflection. In other words, when a sample contains a component that was not present in the training set, or when the sample in any other way (such as its particle size distribution) lies outside the "domain" of the training set, erroneous composition values can result without any indication of the error.

One cure for this problem would be to find a method of identifying "strange" samples using only their **near-infrared** spectra, and indeed, such a method **exists**.²⁴ The use of such a method allows different models to be applied in the determination of the compositions of different samples. This identification method collects reflectance data at n wavelengths, and represents each wavelength as a particular spatial dimension. A sample spectrum taken at n wavelengths can then be represented as a single point in an n -dimensional space. The point is translated from the origin by amounts that correspond to the magnitude of **reflectance** observed at each wavelength. A group of similar samples with similar spectra

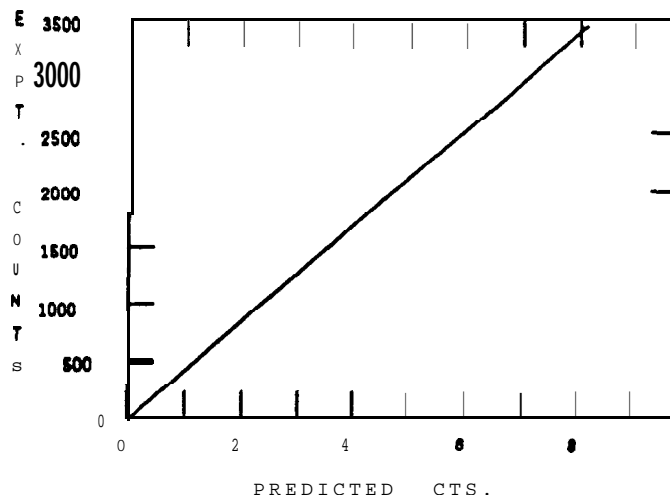


FIG. 8. A QQ plot of the data in Fig. 5 (on the ordinate) vs. the shift function Γ (on the abscissa) when $\alpha_{\text{ref}} = \alpha_{\text{exp}}$.

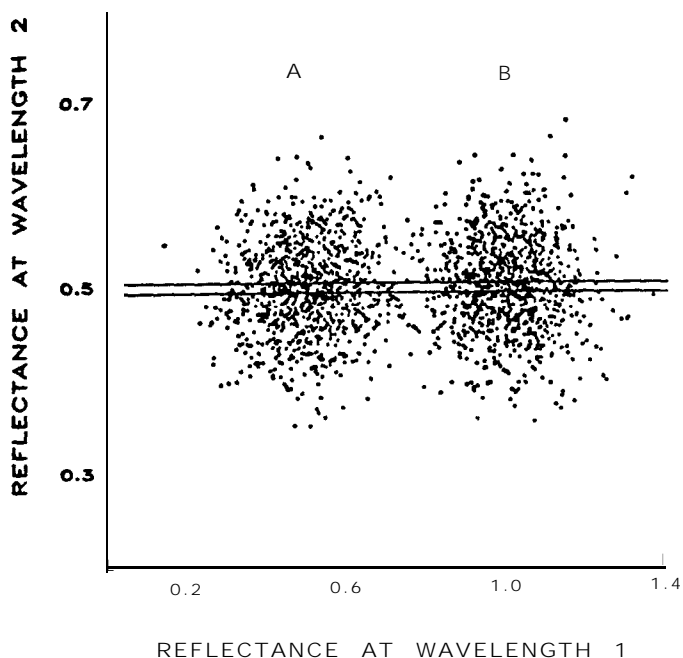


FIG. 9. Two clusters, A and B, represent hypothetical spectral reflectance data collected at the two wavelengths, 1 and 2, for compounds A and B.

appears as a cluster of points at a given location in space (see Fig. 9).

In our laboratory we have investigated the behavior of these points and clusters in hyperspace using quantile analysis. The analyzed distribution (the ECDF) is that of the density of points in a given direction from the center of a cluster, because the clusters themselves are multidimensional, and QQ plots can plot only one univariate distribution against another. Quantiles are used to set confidence limits around the clusters by plotting density in a given direction. Figure 10 shows a quantile plot of the clusters in Fig. 9. This plot was generated along a line (shown in Fig. 9) connecting the centers of the two clusters. The slopes and intercepts of the two lines in this QQ plot define equations for the point-probability density in the direction through the two cluster centers. With this information, one can assign points to a cluster, effectively allowing unknowns to be identified from their near-IR spectra. This method permits samples to be pre-screened before quantitation to ensure that the proper linear model is chosen for the measurement process.²⁵

CONCLUSIONS

Distribution analysis using QQ plots, unlike basic CDF plots, is robust, because judging whether the data points lie on a straight line (and therefore come from a specified parameterized family) is insensitive to the location and scale of the data. Examination of the plots in Figs. 3 and 4 shows that QQ plots are capable of making the tails of distributions more visible. Finally, distribution analysis using QQ plots is more powerful than the classical statistical approach that:

1. proposes a null hypothesis that the distributions are identical;

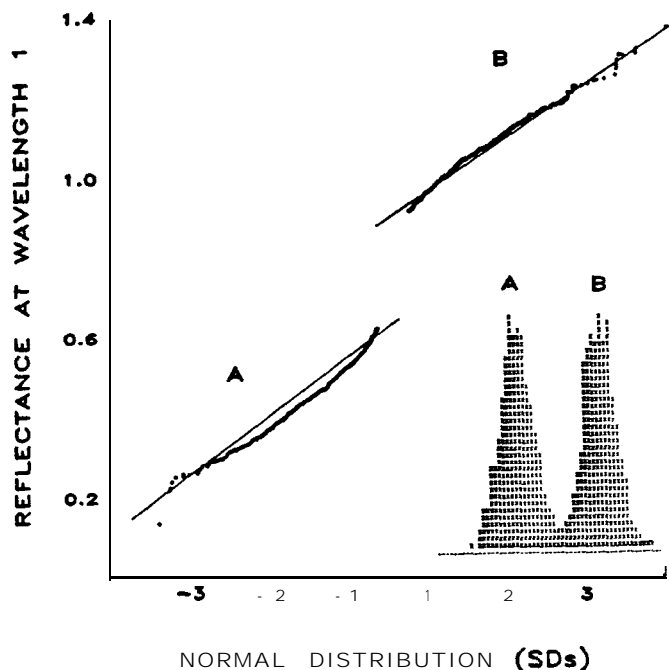


FIG. 10. A quantile plot of clusters in Fig. 5, taken along a line connecting the centers of A and B. The insert shows a histogram of these clusters, taken along the same line.

2. calculates a suitable test statistic from the data (whose distribution is known only if the null hypothesis is true); and
3. fails to accept the null hypothesis if the value of the test statistic falls outside a certain range based upon a previously selected significance level.

If the classical approach fails to accept the null hypothesis, no real indication of what happened is given, and little information is available regarding how best to proceed. Even if the null hypothesis holds, there is no real guarantee that the distributions actually match; each test statistic is sensitive only to certain types of departures from the proposed distribution. In addition, many real cases are borderline, and if a different arbitrary significance level had been selected, the opposite answer would have been obtained.

Quantile plots give a great deal of information about distributions. In a single plot, they can isolate the first three moments of a set of observations: the location, the scale, and the direction of any skew. This information can be analyzed by curve-fitting, modeling, and pattern-recognition techniques. The flexibility of the QQ plotting procedure makes it an obvious first choice in the exploratory analysis of distributions.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation through Grant CHE 87-22639, by the Office of Naval Research, and by the Upjohn Company.

1. R. C. Engstrom, M. Weber, and J. Werth, *Anal. Chem.* **57**,933 (1985).
2. H. N. Cheng, *Anal. Chem.* **54**, 1828 (1982).
3. S. C. Abrahams and J. L. Bernstein, *J. Chem. Phys.* **55**, 3206 (1971).
4. S. C. Abrahams, J. L. Bernstein, and E. T. Keve, *J. Appl. Cryst.* **4**, 284 (1971).

5. M. Seshasayee, *Acta Cryst.* **C39**, 1473 (1983).
 6. A. Parczewski, *Anal. Chim. Acta* 139, 221 (1981).
 7. W. W. Daniel, *Introductory Statistics with Applications* (Houghton Mifflin, Boston, 1977).
 8. J. B. Kennedy and A. M. Neville, *Basic Statistical Methods for Engineers and Scientists* (Harper and Row, New York, 1976).
 9. P. A. Tukey, *Proc. Symp. Appl. Math.* **28**, 9 (1983).
 10. P. A. Tukey, *Proc. Symp. Appl. Math.* **28**, 8 (1983).
 11. N. P. Jewell, *Modern Data Analysis* (Academic Press, New York, 1982).
 12. S. M. Sharma and S. K. Sikka, *Phil. Mag. B* 45, 317 (1982).
 13. K. C. Yang and R. Hogg, *Anal. Chem.* 51, 758 (1979).
 14. L. J. Nagels, W. L. Creten, and P. M. Vanpeperstraete, *Anal. Chem.* 55, 216 (1983).
 15. V. J. Nair, *Amer. Statist. Assoc.* 79, 823 (1984).
 16. S. W. Looney and T. R. Gullledge, *Amer. Stat.* 79, 75 (1985).
 17. N. L. Morgenstern and M. A. Shearn, *Am. J. Clin. Pathol.* 76, 211 (1981).
 18. M. D. Knittel, C. H. Black, W. E. Sandine, and D. K. Fraser, *Can. J. Microbiol.* 14, 239 (1968).
 19. S. C. Abraham and E. T. Keve, *Acta Cryst.* **A27**, 157 (1971).
 20. S. W. Looney and T. R. Gullledge, *J. Statist. Comput. Simul.* 20, 115 (1984).
 21. M. Huang and G. M. Hieftje, *Spectrochimica Acta* **40B**, 1387 (1985).
 22. E. E. Salpeter, *Phys. Rev.* **120**, 1528 (1960).
 23. D. L. Wetzel, *Anal. Chem.* 55, 1165 (1983).
 24. R. A. Lodder, M. Selby, G. M. Hieftje, *Anal. Chem.* **59**, 1921 (1987).
 25. R. A. Lodder and G. M. Hieftje, *Appl. Spectrosc.* **42**, 1500 (1988).
-