Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 509 (2004) 197–208
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Detection limits and goodness-of-fit measures for the two-component model of chemical analytical error

Machelle D. Wilson [a,*], David M. Rocke [b], Blythe Durbin [b], Henry D. Kahn [c]

[a] *Savannah River Ecology Laboratory, The University of Georgia, P.O. Drawer E, Aiken, SC 29802, USA*
[b] *University of California, Davis, CA, USA*
[c] *United States Environmental Protection Agency, USA*

Received 8 July 2003; received in revised form 12 December 2003; accepted 19 December 2003

Abstract

The utility of analytical chemistry measurements in most applications is dependent on an assessment of measurement error. This paper demonstrates the use of a two-component error model in setting limits of detection and related concepts and introduces two goodness-of-fit statistics for assessing the appropriateness of the model for the data at hand. The model is applicable to analytical methods in which high concentrations are measured with approximately constant relative standard deviation. At low levels, the relative standard deviation cannot stay constant, since this implies vanishingly small absolute standard deviation. The two-component model has approximately constant standard deviation near zero concentration, and approximately constant relative standard deviation at high concentrations, a pattern that is frequently observed in practice. Here we discuss several important applications of the model to environmental monitoring and also introduce two goodness-of-fit statistics, to ascertain whether the data exhibit the error structure assumed by the model, as well as to look for problems with experimental design.

Keywords: Chemical analytical error; Limits of detection; Two-component error model; Goodness-of-fit

1. Introduction

Limitations of the analytical methodology used to measure the concentration of toxic substances in the environment have had an important policy role in regulation. It is difficult to regulate emissions of toxic substances at levels below what can be reliably measured, and a definition of the level of reliable measurement is therefore crucial to policy making. In this paper, we discuss the implications of a model for measurement error for these policy issues.

It has been observed from long experience that the measurement error of an analytical method, for example atomic absorption spectroscopy, is of two types. Over a range of concentrations near zero, the measurement error is seen to be constant. Over ranges of higher concentration, the measurement error is observed to be proportional to the concentration of the analyte [1,2]. This poses some difficulty in estimating the overall precision of an analytical method for data that span the "gray area" where a transition occurs between near zero concentrations and quantifiable amounts. If a model assuming a constant error is used, there is an implicit assumption that the absolute size of the error is unrelated to the concentration of the analyte. This assumption is not supported by empirical observation of behavior at the higher levels. If a model assuming a proportional error is used, then there is an implicit assumption that the measurement error becomes vanishingly small as the concentration approaches zero. This assumption is also contrary to experience of behavior at the lower levels. Since environmental monitoring data often fall into this gray area, understanding measurement error in this region is of considerable importance.

The model presented here was first proposed by Rocke and Lorenzato [3], where some of the technical background to this model can be found. The model resolves the difficulties discussed above by incorporating both types of error that are observed in practice into a single model. This model provides an obvious advantage over existing models by describing the precision of measurements across the entire usable range. We will present two examples in detail—one of Ni by ICP/MS and one of proprionitrile by GC/MS—numerical illustrations using Zn and Ag by ICP/MS, and summary results for a larger group of organics by GC/MS and metals by ICP/MS. These examples support the validity and

* Corresponding author.

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2003.12.047

advantages of this two-component model. We also discuss the application of the model to detection limits, quantification limits, sample size calculations, and the construction of confidence intervals. We introduce two goodness-of-fit statistics and estimate their distributions using the parametric bootstrap. Some related work with a different point of view on some of the policy issues can be found in [4–6].

It should perhaps be noted here that this model deals with variation in the calibration curve produced by the analytical instrument only. It does not deal with other sources of variation, such as the inter-laboratory variation, human-error, or error in identification of peak area in gas chromatography/mass spectroscopy. Identifying a signal as a peak is a non-trivial task and the inherent variation produced by this difficulty is beyond the scope of this paper. For further readings in this area, see [7].

2. The model

Most measurement technologies require a calibration curve, often linear, to estimate the actual concentration of an analyte in a sample for a given response. We can incorporate into the linear calibration model the two types of errors that are observed in most analyses. The two-component model is

$$y = \alpha + \beta\mu\, e^{\eta} + \epsilon \tag{1}$$

where y is the response (such as peak area) at concentration μ, α and β are the parameters of the calibration curve, $\eta \sim N(0, \sigma_\eta)$ and $\epsilon \sim N(0, \sigma_\epsilon)$. Here, η represents the proportional error that always exists, but dominates at concentrations significantly above zero, and ϵ represents the additive error that always exists but dominates mainly for near zero concentrations. This two-component model approximates a constant standard deviation for very low concentrations and approximates a constant relative standard deviation (RSD) for higher concentrations. Note that y is the response of the measuring apparatus, for example peak area. In order to obtain the estimated concentration, we perform the back calculation

$$\hat{\mu} = \frac{y - \hat{\alpha}}{\hat{\beta}} \tag{2}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β, respectively, in the model (1).

Under this model, the variance of the response y at concentration μ when α and β are known is given by

$$\mathrm{Var}\{y\} = \mu^2\beta^2\, e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\epsilon^2 \tag{3}$$

This calculation relies on the assumption that α and β are known, which is perhaps an unusual circumstance. It is safe to consider these parameters known when the estimates are based on large amounts of available historical data where there is little variation in the observed calibration line over time. When the regression and variance parameters are unknown, these quantities can be estimated by simply substituting the estimates from the algorithms. This step is justified in application, since, as the bootstrap results show, the variance of the parameter estimates for maximum likelihood estimation (MLE) is quite small [3].

Two derived quantities will be useful in interpretation of the results. $S_\epsilon = \sigma_\epsilon/\beta$ represents the standard deviation of $\hat{\mu}$ at low levels.[1] $S_\eta = \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)}$ is the RSD of $\hat{\mu}$ for high levels. For values of σ_η appropriate for analytical technologies (say not more than 0.3), S_η is very near σ_η. For example, if $\sigma_\eta = 0.1$, then $S_\eta = 0.1008$ and if $\sigma_\eta = 0.3$, then $S_\eta = 0.32$.

Using these derived quantities, we can represent the variance of y as

$$\mathrm{Var}\{y\} = \mu^2\beta^2 S_\eta^2 + \sigma_\epsilon^2 \tag{4}$$

and for the estimated concentration,

$$\mathrm{Var}\{\hat{\mu}\} = \mu^2 S_\eta^2 + S_\epsilon^2 \tag{5}$$

The usefulness of the two-component error model is clear when compared to using only the relative standard deviation, which is defined to be equal to the standard deviation of the estimated concentration divided by the concentration [8]. For the two-component model, we have

$$\mathrm{RSD}\{\hat{\mu}\} = \sqrt{S_\eta^2 + \frac{S_\epsilon^2}{\mu^2}} \tag{6}$$

If the error structure is described only in terms of RSD, we see that measurements at high concentrations have a nearly constant RSD, whereas small concentrations have an increasing RSD that tends to infinity as the concentration approaches zero, which is not observed in practice. Use of RSD alone to characterize measurement error in the low concentration region can cause difficulties when attempting to make decisions regarding detection and quantification. The two-component model allows for a more reasonable estimation of error near zero, and hence more reasonable criteria for setting detection limits and critical levels.

Likewise, using the simple model, i.e. assuming that the standard error remains constant across all concentrations, results in grossly inaccurate estimates of error at higher concentrations. For example, assume that the standard deviation of blanks is 29 g/l, but that $\sigma_\eta = 0.3$, a very high coefficient of variation that nevertheless often occurs with some technologies, and that $\beta = 1$. Suppose we have an estimated concentration of 600 g/l and would like to calculate confidence intervals around that estimate. Using the simple model, we would estimate the standard deviation to be that of the blanks, or 29 g/l. However, since $\sigma_\eta = 0.3$, the true standard deviation is $\sqrt{29^2 + 600^2 S_\eta^2} = 182$ g/l, an increase of the variance of over six times the variance at zero. Poor

[1] Here we assume that α and β are well enough estimated that they can be treated as known.

estimation of error at a given concentration will heavily affect the estimation of limits of detection, and hence regulatory practices.

3. Estimation

The parameters in the two-component model can be estimated in a number of ways. The standard deviation σ_ϵ of the low level measurements can be estimated from replicate blanks, for example by those routinely included with batches of samples. If this number is stable, it can also be estimated from routine QC data so long as measurements at zero concentration are replicated. The parameter σ_η can be likewise estimated from the standard deviation of the logarithm of high level measurements. The calibration curve can then be estimated using weighted least squares, weighting each point by the inverse estimated variance (using (4) with estimates of the parameters inserted). It is also possible to estimate all four parameters simultaneously using weighted least squares, although our experience is that this estimation method is often not very stable and can lead to non-convergence or impossible estimates (such as negative variances).

The most effective estimation method is maximum likelihood estimation, as described in [3]. A computer program that solves for the maximum likelihood estimates for α, β, σ_ϵ, σ_η is available at http://www.cipic.ucdavis.edu/~dmrocke.

Two example data sets and the results from the maximum likelihood algorithm will be shown here, as well as summary results for several other data sets.

4. Limits of detection

In this section we describe some applications of the two-component model. Special emphasis is given to detection and measurement of toxins at very low levels.

4.1. Critical levels

Detection refers to the capability of an analytical measurement process to indicate the presence of an analyte. This requires an agreed upon procedure for determining whether or not a given measurement result conclusively establishes that the analyte is present in the sample [9]. In practice, this means that the investigators establish a numerical value such that a result greater than this value is extremely unlikely to occur if the true concentration is in fact zero, while a result lower than this value indicates that the true concentration in the sample is either zero, or is too low to detect (with certainty) with the technology in use. The measurement error that exists in any technology leads to this inability to detect concentrations below a certain level. The critical level is defined by the International Union of Pure and Applied Chemistry (IUPAC) to be the value, L_C, such that the probability of a measurement exceeding this value will be very small, say 0.01, when the true concentration in the sample is zero [10]. That is, samples that do not contain the analyte are very unlikely to generate measurement results that exceed L_C. Note that the critical level is defined at first in the units of the measurement technology (e.g., peak area) not in units of concentration. Of course, we can also express the critical level in units of concentration by taking the critical level (in measured units) and dividing by the calibration slope β. Since the critical level is the point at which the detection decision is made, it has been called by some authors a detection limit, but it should be noted that it is distinct from the IUPAC definition of the limit of detection. Limits of detection will be discussed later in this paper.

Under the assumption of normality, the value of L_C may be calculated as follows: assume that σ_ϵ, the standard deviation of the response at $\mu = 0$, is known and that we require 99% confidence in our statement that the analyte is present. Then the one-sided 99% confidence level is represented by $L_C = \alpha + z_0\sigma_\epsilon$, where z_0 is the z-value corresponding to the 99th percentile of the standard normal distribution (i.e., $z_0 = 2.326$). To find the critical level for any level of confidence, simply find the appropriate one-sided z-value, then multiply by the standard deviation of the blanks and add this to the mean value of the blanks, i.e.,

$$L_C = \alpha + z_0\sigma_\epsilon \tag{7}$$

in units of the response of

$$L_C = z_0 S_\epsilon \tag{8}$$

in units of concentration. Generally, the mean and standard deviation of the blanks will be well enough known from experience to use this method. If these are estimated from data, then a t-value (from the t-distribution), with the appropriate degrees of freedom, is substituted for the z-value. The advantage of the two-component error model is that an estimate for σ_ϵ with desirable statistical properties can be obtained from data that span a range of concentrations, resulting in greater accuracy from a given amount of data. Specifically, the two-component model accounts for the inherent error structure of the entire data set. Once we have achieved adequate goodness-of-fit by considering the error structure as a whole, the parameter estimates provide us with more accurate estimates of error at the concentrations we are interested in, here, concentrations near the limit of detection.

For our example zinc data, we have $\sigma_\epsilon = 204$ in units of peak area and $S_\epsilon = 28.9$ g/l. If we use a 99% confidence level, the normal percentage point is 2.326, so that L_C in units of peak area is $490 + (2.326)(204) = 965$ and in units of concentration is $(2.326)(28.9) = 67.2$ g/l.

Note that a measured value below L_C does not establish that the analyte is absent, only that it has not been shown conclusively to be present. This means that the value should be reported as measured, together with the standard deviation at the measurement value. Cases in which limitations of the instrument itself prevent reporting a value (e.g. as

is the case with some spectroscopic measurements) are an obvious exception. In other cases, censoring of values below L_C may be required as a matter of policy by regulatory agencies. Such practices result in data sets with reduced utility from the loss of information, which makes tracking of trends, monitoring of laboratory quality, summarization of data, and other data analysis all more difficult. More importantly, this censoring needlessly prevents investigators from being able to reach probabilistically quantified conclusions about the true presence of an analyte. Sometimes more or less sophisticated methods can be used to cope with these censored data [11–13], but simple reporting of measured values would allow use of basic, easily understood methodology instead of these complex techniques.

4.2. Minimum detectable value

The limit of detection or minimum detectable value is the true concentration, L_D, of an analyte that will, with high confidence, produce a measured value above the critical level. For example, if the concentration L_D is chosen for laboratory QC, it should be detected (measured above the critical level L_C) almost all of the time [14]. Although L_C can be given either in the units of the measurement technology or in units of concentration, L_D is purely in units of concentration. Conceptually, L_C is determined so that the desired confidence level of the test that the true concentration is zero is met, and L_D is determined so that the desired statistical power is obtained. It usually cannot be safely assumed that the standard deviation at the detection limit is the same as the standard deviation of a blank, so reliable estimates of variance at any specified concentration are necessary for reliable determination of the minimum detectable value.

We can find a good estimate of L_D by noting that the level is low enough that a normal approximation is appropriate (at high levels, the distribution is essentially log normal). We treat y as being normally distributed with mean $\alpha + \beta\mu$ and with variance given by (4) and solve the resulting equation. When the standard deviation is estimated, the following is analogous, but we use the appropriate quantile from the Student's t distribution rather than that from the normal. Recall that, from (5),

$$\mathrm{Var}\{\hat{\mu}\} = \mu^2 S_\eta^2 + S_\epsilon^2 \tag{9}$$

so that L_D is the solution to the equation

$$L_D = z_0 S_\epsilon + z_1 \sqrt{\mathrm{Var}\{L_D\}} \tag{10}$$

$$L_D = z_0 S_\epsilon + z_1 \sqrt{L_D^2 S_\eta^2 + S_\epsilon^2}, \tag{11}$$

where z_0 is the percentile of the standard normal distribution corresponding to the desired confidence level for the critical level, and z_1 the percentile corresponding to the desired confidence level for the minimum detectable value. So long as the variance of y does not increase too rapidly with μ, a closed form solution is possible. The required condition is

$$S_\eta < \frac{1}{z_1}. \tag{12}$$

For details and proof, see Appendix A. Note that if this condition is not satisfied, then there is no solution at all. The closed form solution derived then is

$$L_D = \frac{S_\epsilon \left[z_0 + \sqrt{z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)} \right]}{1 - z_1^2 S_\eta^2} \tag{13}$$

When $z_0 = z_1 = z$, we have the particularly simple form

$$L_D = \frac{2 z S_\epsilon}{1 - z^2 S_\eta^2} \tag{14}$$

Then an estimated L_D is found by simply substituting the sample variance estimates into the above equation.

For our example zinc data, we have $S_\epsilon = 28.9$ and $S_\eta = 0.0390$. If we use $z_0 = z_1 = 2.326$, corresponding to 99% confidence, we have a minimum detectable value of

$$\frac{(2)(2.326)(28.9)}{1 - (2.326)^2 (.0390)^2} = 135\,\mathrm{g/l} \tag{15}$$

One important use of L_D is to assess and monitor the performance of a laboratory. If samples are spiked at a concentration of L_D, then almost all of the time the resulting responses should exceed the critical value L_C. Such trials can be run periodically to monitor the ability of the laboratory to detect analytes up to specifications. Another important use is to determine what concentrations in the field can reliably be detected with a given technology. If concentrations below the minimum detectable value are important to detect, consideration should be given to the use of better technology or replicate measurements. The minimum detectable value should never be used to assess a measured value to decide if it should be reported or censored. It should only be used for planning purposes or for quality assurance.

4.3. Quantification limit

By a quantification limit, some authors have meant the lowest level at which the quantitative assessment is sufficiently accurate for practical use. Since the standard deviation at low levels is actually smaller than that at high levels, the most precise measurements, in terms of standard deviation, are actually those for the lowest level of the analyte. A definition with some practical utility is the true concentration at which the relative standard deviation falls to a specified level [6,15,16]. However, measurement of some analytes at an arbitrarily low RSD, such as 10% may not be possible. The model allows for evaluation of each case in terms of what RSD is feasible. The RSD at 0 is automatically infinite, no matter how accurate the measurement process is, and the RSD at high levels is σ_η so it is meaningful to define the quantification limit as the level at which the RSD falls to a specified multiple of σ_η, say $2\sigma_\eta$. It makes little sense

to specify a particular arbitrary RSD, such as 20%, since the limit of quantification would then be undefined for any measurement process with $\sigma_\eta > 0.20$. This is easily seen by substituting in 0.2 in the example given earlier. If it is desirable to make this computation for a particular process, this can easily be done using the model presented here. Let R be the desired RSD, e.g. 10%. Then the equation

$$\frac{\sqrt{V(L_Q)}}{L_Q} = R \tag{16}$$

where $V(L_Q) = L_Q^2 S_\eta^2 + S_\epsilon^2$ has solution

$$L_Q = \sqrt{\frac{S_\epsilon^2}{R^2 - S_\eta^2}} \tag{17}$$

whenever $R > S_\eta$. No real solution is possible when $R \leq S_\eta$. This is readily apparent once (16) is re-written as a quadratic equation in L_Q.

For our zinc example, $S_\epsilon = 28.9$ ppt $S_\eta = 0.0390$. If the desired RSD is 10%, then

$$L_Q = \frac{28.9}{\sqrt{(0.1)^2 - (0.0390)^2}} = 314 \text{ ppt} \tag{18}$$

Compare this to the 99% confidence critical level of 67.2 ppt and the minimum detectable value of 135 ppt. The limit of quantification is somewhat arbitrary by comparison. While the critical level and the minimum detectable value are quantified using standard normal probability theory, the limit of quantification is set arbitrarily by the investigator. For example, if the target RSD is chosen to be 15%, rather than 10% then the quantification limit is 200 ppt instead of 314 ppt. For analytes that are toxic at very low levels, this arbitrary choice may have rather severe consequences. As in the case of the minimum detectable value, the limit of quantification has no use in interpreting measurements that have already occurred. The estimated concentration along with a measure of the uncertainty of the measurement convey all of the necessary information.

5. Goodness-of-fit tests

As the two-component model is relatively new, there are no existing goodness-of-fit tests available for evaluating the conformity of the data to the model, or the efficacy of the various algorithms used to fit the data to the model. Hence, we need goodness-of-fit statistics to help determine if the data indeed fit the model, and as a way of comparing different methods of estimation.

5.1. Error structure

Since estimates for α and β are easily and reliably obtained using standard weighted linear regression techniques and since the variance estimates are computationally intensive and crucial to important applications of the model, we focus on comparing the model predicted variance σ_i^2 at each concentration to the mean square deviation $\tilde{s}_i^2$ from the calibration curve. Hence we propose the following goodness-of-fit statistic for testing the appropriateness of the modeled error structure for the data. Values of $\sigma_i^2/\tilde{s}_i^2$ close to 1 (or values of $\log \sigma_i^2/\tilde{s}_i^2$ close to 0) indicate a good fit.

For the maximum likelihood estimation method, the predicted variance at a given concentration was estimated using the formula

$$\sigma_i^2 = \sigma_\epsilon^2 + \beta^2 \mu_i^2 \, \mathrm{e}^{\sigma_\eta^2} (\mathrm{e}^{\sigma_\eta^2} - 1), \tag{19}$$

where μ_i is the concentration.

Suppose, for example, that one wishes to calculate the ratio $\sigma_i^2/\tilde{s}_i^2$ for the maximum likelihood method at a concentration of $\mu = 100.0$. Suppose, furthermore, that we have the results $\hat{\alpha} = 114.80$, $\hat{\beta} = 11.586$, $\hat{\sigma}_\eta = 0.028424$, and $\hat{\sigma}_\epsilon = 10.525745$, and that there are $r = 5$ replicates for $\mu = 100.0$ with the values $y_{100,1} = 1286$, $y_{100,2} = 1239$, $y_{100,3} = 1273$, $y_{100,4} = 1177$, $y_{100,5} = 1306$. We calculate the predicted value

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\mu_i$$

or

$$\hat{y}_i = 114.80 + 11.586(100.0)$$

giving a predicted response of 1273.2. We then calculate $\tilde{s}_i^2$ using the formula

$$\tilde{s}_i^2 = \frac{1}{r} \sum_{j=1}^{r} (y_{i,j} - \hat{y}_i)^2 \tag{20}$$

and obtain $\tilde{s}_{100}^2 = 1413.7$. Calculating σ_{100}^2 from (19), we obtain $\sigma_{100}^2 = 1196.6$ and $\sigma_{100}^2/\tilde{s}_{100}^2 = 0.84648$, indicating a fairly good fit at this concentration.

The goodness-of-fit statistic for an entire data set is

$$T_{\mathrm{gf}} = \ln \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_i^2}{\tilde{s}_i^2} \right] \tag{21}$$

where n is the number of concentrations. A value near zero for T_{gf} indicates a good fit.

5.2. Good experimental design

When running standards to obtain a calibration curve, the best experimental design is to randomize the replicates, so that the replicates at a given concentration are not all run sequentially. The reason for this is that when the replicates at a given concentration are all run sequentially, values will tend to be closer together than if the replicates are randomized among the other concentrations. This often results in a calibration curve that runs through the data in such a way that at some concentrations all the data are bunched above the curve and at others all the data are bunched below the curve. A graph of the data and the calibration curve that

shows this behavior indicates poor experimental design. An extreme example of this phenomenon is shown in Figs. 4 and 5. If the replicates are correctly randomized, the mean square deviation away from the mean will be very close to the mean square deviation away from the calibration curve. Thus, to detect possible poor experimental design and distinguish it from true lack-of-fit in the inherent error structure of the analytical instrument, we propose the statistic

$$S_{\mathrm{gf}} = \frac{1}{n}\left[\sum_{i=1}^{n} \ln\left\{\frac{\hat{s}_i^2}{\tilde{s}_i^2}\right\}\right] \tag{22}$$

where $\hat{s}_i^2$ is the unbiased sample variance at a given concentration and $\tilde{s}_i^2$ the mean square deviation away from the calibration curve, as above. Values of S_{gf} close to zero indicate good randomization.

It should be noted that data with a non-linear calibration curve will produce a high absolute value for S_{gf} if the data are fit to a linear model. Basic visual diagnostics easily reveal whether the lack-of-fit is caused by poor experimental design or a non-linearity in the data. If the true calibration curve is non-linear, the data will lie along a quadratic or exponential curve. If the lack-of-fit is due to poor experimental design, the data will approximately follow the linear calibration curve, but will bunch above and below it randomly across concentrations. If the data follow an exponential curve, then an exponential calibration model, rather than a linear one, should be fit.

6. Methods: the parametric bootstrap

Because the exact distribution of any goodness-of-fit statistic for this model would be difficult to derive, we use the parametric bootstrap to obtain an estimate of the null distribution. We can also use the bootstrap to obtain estimated distributions of the parameter estimates, which are helpful for assessing the efficacy of the estimation algorithm.

The parametric bootstrap [17] is performed by first obtaining the parameter estimates from the algorithms. These parameter estimates are used as parameter values to generate data sets that we know fit the model. This procedure is repeated a large number of times (about 1000) for each analyte. These simulated data sets are then run through the estimation routines and parameter estimates for each parameter for all 1000 data sets for each analyte are obtained.

For example, the parameter estimates using the maximum likelihood routine for the Ag data set (results not shown) were: $\hat{\alpha} = 21.4022$, $\hat{\beta} = 27.4918$, $\hat{\sigma}_\epsilon = 43.4769$, and $\hat{\sigma}_\eta = 0.0235$. We generate measurement errors $\epsilon \sim N(0, \hat{\sigma}_\epsilon^2)$ and $\eta \sim N(0, \hat{\sigma}_\eta^2)$. We can then generate an artificial data set that follows the fitted model by setting $y = \hat{\alpha} + \hat{\beta}\mu\, \mathrm{e}^{\eta} + \epsilon$ using the same value of μ as in the original data set. This procedure is repeated 1000 times, generating 1000 data sets. These simulated data sets are then run through the estimation routine and the parameter estimates and goodness-of-fit statistics saved in a file. By plotting histograms of the 1000 parameter estimates and the statistics, and ranking their values, we can obtain an estimate of the distribution and its quantiles.

A 95% bootstrap confidence interval can be obtained by sorting the vectors of parameter estimates. The 95% confidence interval is then the interval between the 25th and 975th data points in the sorted data vector. If the estimation routine is working well, and the data fit the model, all the histograms of the 1000 parameter estimates will be centered at the true value and approximately bell-shaped about this value. For both the goodness-of-fit statistic and the experimental design statistic, the value obtained from the real data set will lie within the 95% bootstrap confidence interval if the data fit the model and the experimental design is optimal.

7. Results

Here we describe the application of the two-component model to two data sets. In one data set, EPA method 1638 (metals by ICPMS) was used in separate analyses to evaluate spiked samples from 0 g/l to a high level (up to 85,000 g/l) depending on the analyte. Twenty-three metals were analyzed, with 8–11 replicates per concentration of each metal. In the second data set, EPA method 5031 (volatile organics by GC/MS) was used to analyze spiked samples for 15 volatile organics, at nine spike levels from 0 to 3000 μg/l, with 4–8 replicates per concentration level.

We give detailed results in this section for one analyte from each data set. The other analytes yielded similar results, and the space required to provide detailed examples for each analyte would be prohibitively large.

7.1. T_{gf}

A scatterplot of T_{gf} for all analytes is shown in Fig. 1. Histograms of the bootstrapped T_{gf} for the example data sets show no or very small skewness for the MLE results (not shown), suggesting little or no bias. A slight skewness was observed in the MLE results, but this is not unexpected, since this statistic is not expected to be exactly normal. Values for T_{gf} which indicate lack-of-fit will vary depending on the error structure of each analyte. The statistical significance of each of the values shown here would have to be generated by the bootstrap, but it appears that the largest three values are inconsistent with the null hypothesis while the others are consistent.

7.2. S_{gf}

We calculated the experimental design statistic, S_{gf} by MLE for each data set. A scatterplot of S_{gf} for all data sets is shown in Fig. 2. As can be seen, poor experimental design is a common problem, with many data sets showing an estimated goodness-of-fit statistic well above zero. We chose nickel as a striking example of this phenomenon.

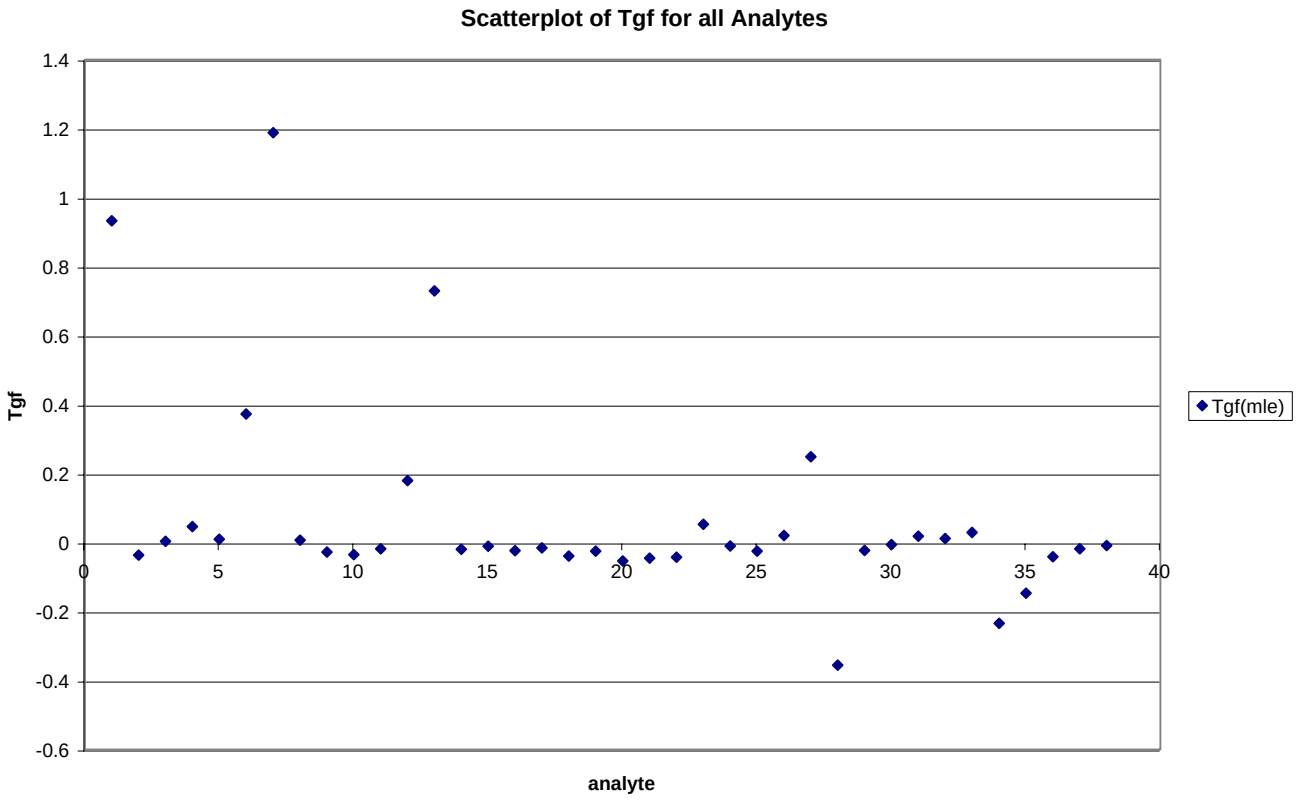

Fig. 1. Scatterplot of T_{gf} for all analytes. Values of T_{gf} near zero show goodness-of-fit.

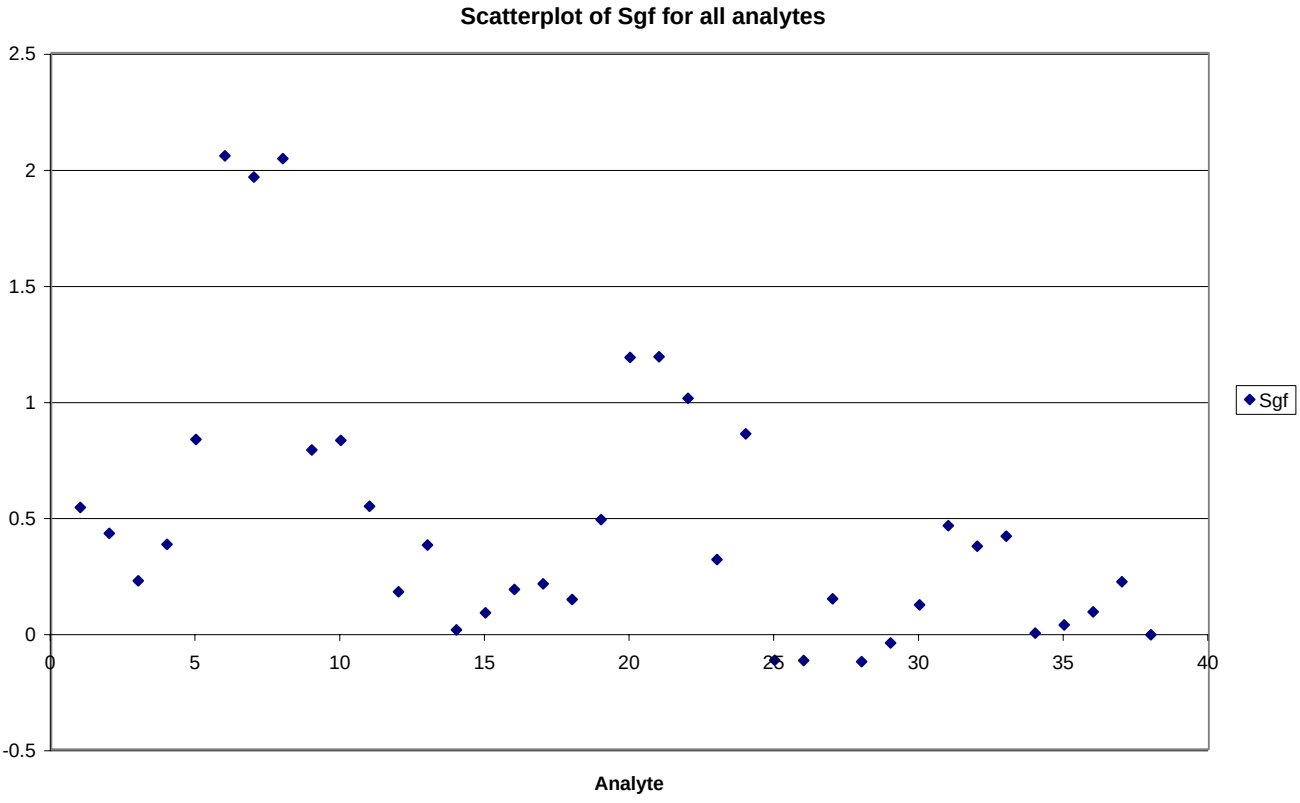

Fig. 2. Scatterplot of S_{gf} for all analytes. Values of S_{gf} near zero show goodness-of-fit.

It is important to point out that since the sum of the squared deviations away from the mean minimizes the sum of the squared deviations away from all possible points, S_{gf} is expected to be positive. However, we are using the unbiased estimate for the sample variance about the mean, i.e., we divide by $n - 1$, but are using the mean square error away from the calibration curve, i.e., we divide by n. Hence, in data sets where the sum of squared deviations about the mean is quite close to the sum of squared deviations away from the calibration curve, we get values for S_{gf} that are slightly negative. This occurs in the acetaldehyde, acetone, and acrylonitrile, acrolein, and proprionitrile data sets.

7.3. Example data sets

Here we show the results for proprionitrile and nickel and the parametric bootstrap analyses. Proprionitrile is an example of a data set that shows both excellent fit and excellent experimental design. The nickel data set shows extremely poor experimental design and hence lack-of-fit as well. Also shown are bootstrap results for the detection decision quantities for proprionitrile.

7.3.1. Proprionitrile

Proprionitrile is an example of a data set that shows both excellent fit and excellent experimental design. In Fig. 3, we see that the observed response scatters nicely around the predicted response. Table 1 shows bootstrap results for the parameter estimates for α, β, σ_ϵ, σ_η the decision quantities L_C and L_D and the goodness-of-fit statistics. Note that the true value is well within the bootstrap confidence interval for all parameters and statistics. The two goodness-of-fit statistics show no lack-of-fit for this data set.

Table 1
Bootstrap results for proprionitrile parameters

Parameter	True value	95% BS confidence interval
α	3.05	(1.69, 4.47)
β	0.891	(0.875, 0.909)
σ_ϵ	3.68	(1.97, 5.35)
σ_η	0.0508	(0.0175, 0.0811)
T_{gf}	−0.00863	(−0.557, 1.32)
S_{gf}	0.00674	(−0.159, 0.346)
L_C	6.81	(3.61, 9.86)
L_D	13.71	(7.34, 19.9)

7.3.2. Nickel

Fig. 4 shows the observed and predicted response for the nickel data set. Note the erratic behavior of the observed response around the calibration line. Fig. 5 highlights this erratic behavior at low concentrations, which are difficult to observe for the full data set, due to scale. The phenomenon associated with poor experimental design is clear. The T_{gf} statistic lies well outside the upper bound of the 95%

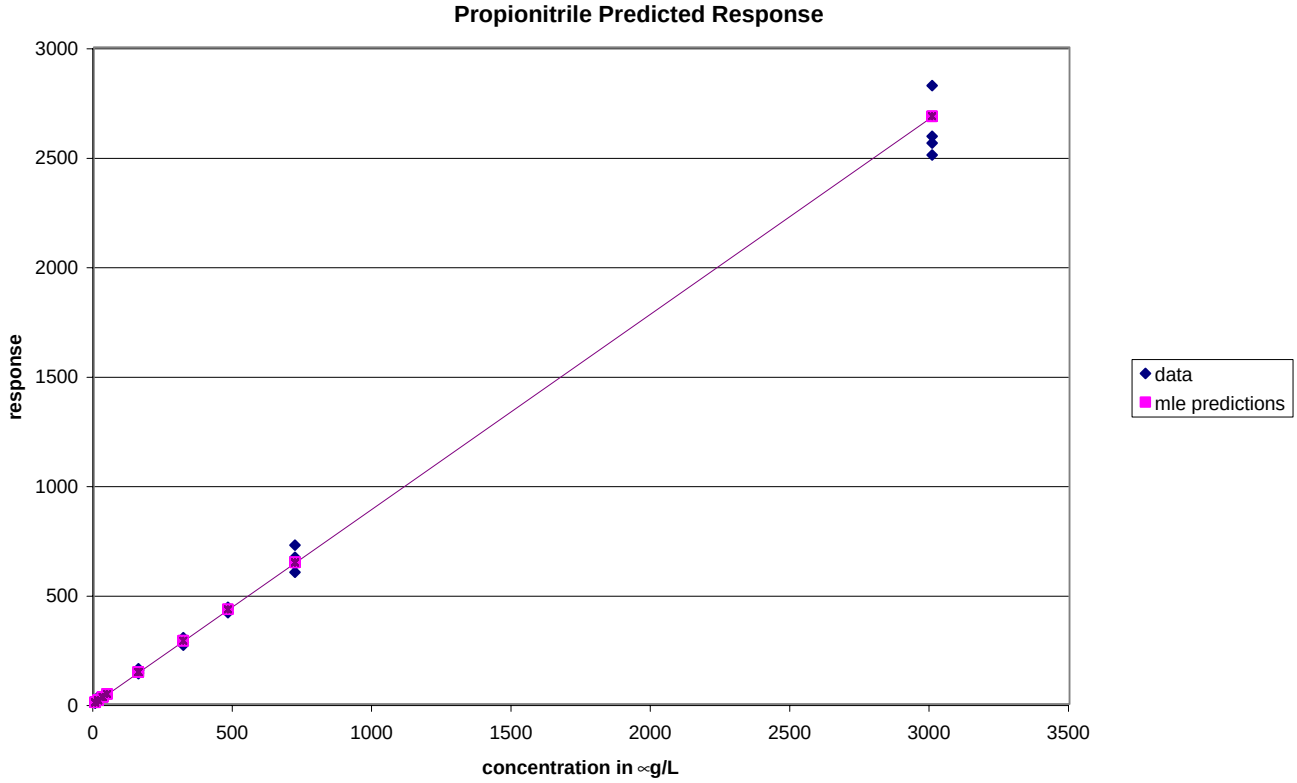

Fig. 3. Proprionitrile predicted response. Estimated calibration line for proprionitrile using maximum likelihood. Note the increasing variance at high concentrations and the near constant variance near zero.

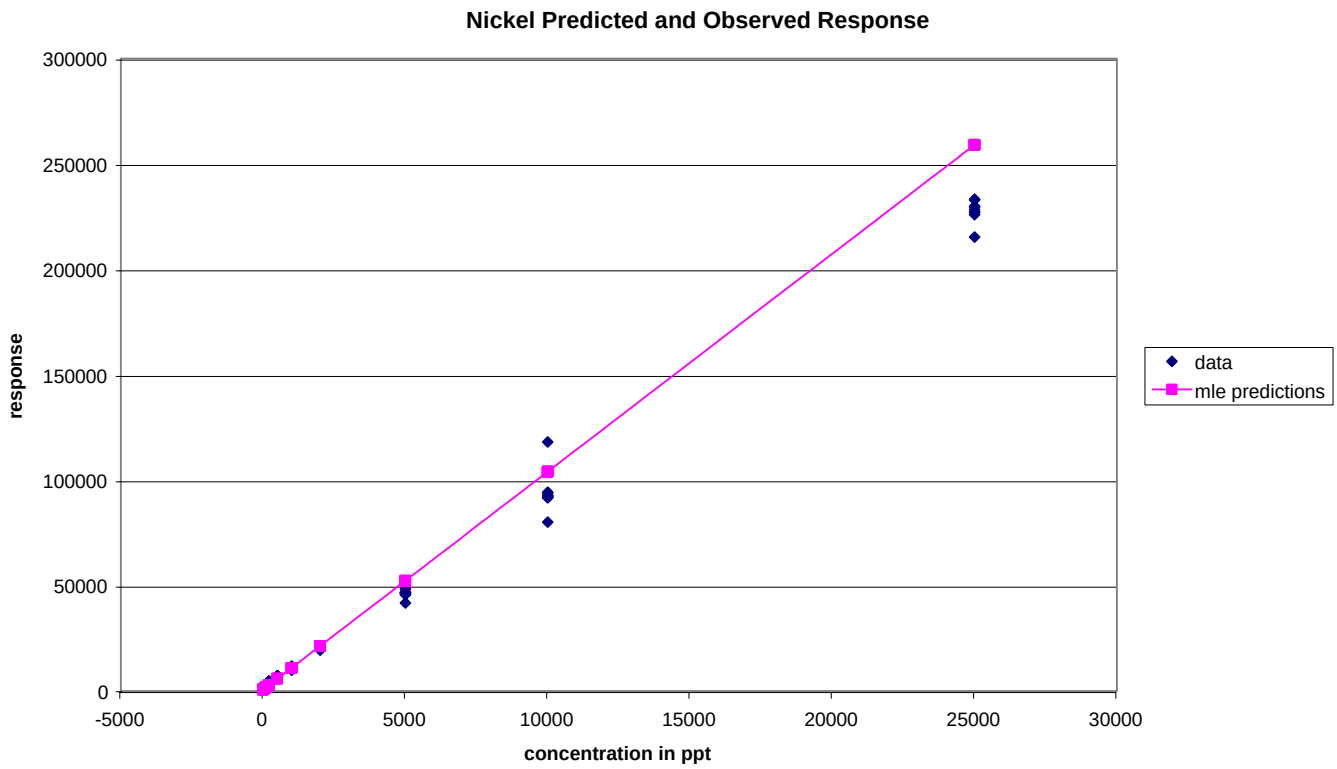

Fig. 4. Nickel predicted and observed response. Estimated calibration line for nickel using maximum likelihood. Note the variance structure, indicating two components of variation and poor experimental design.

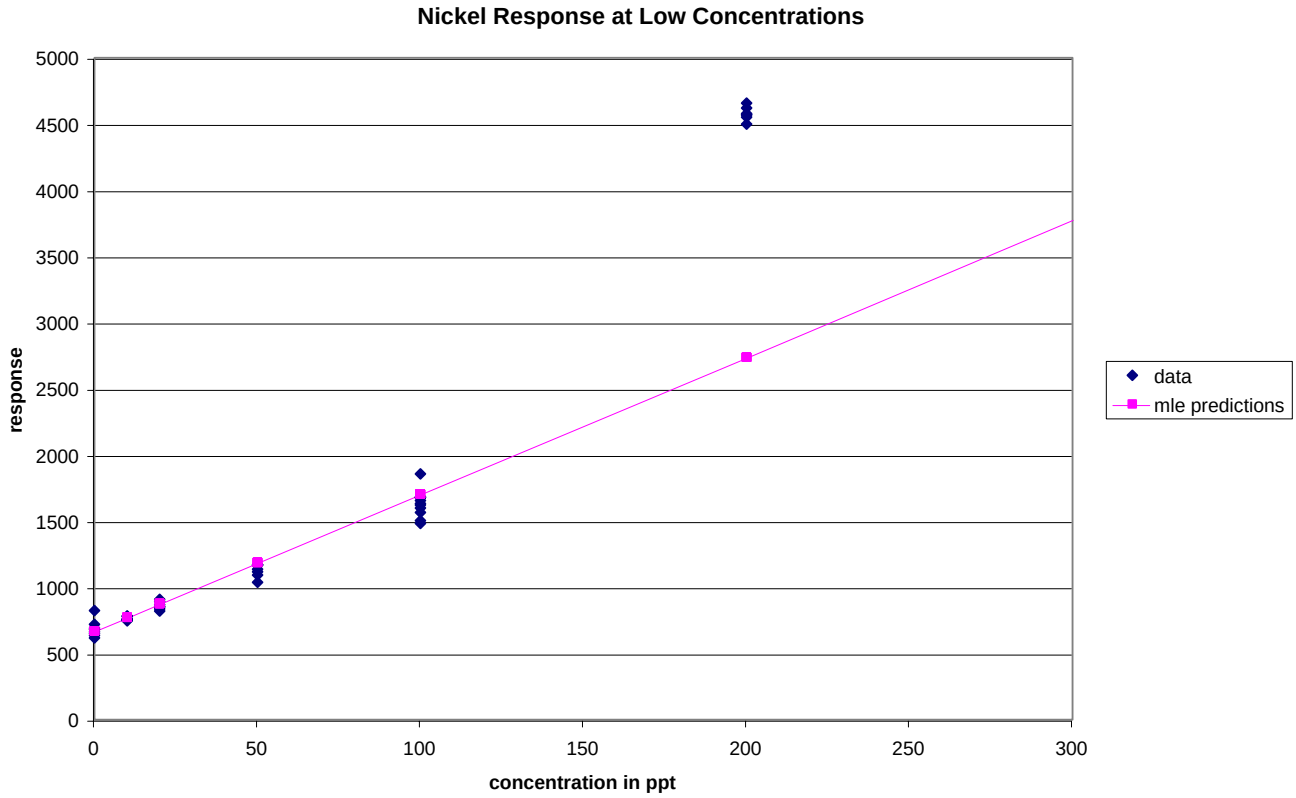

Fig. 5. Nickel response at low concentrations. This figure shows more clearly the erratic nature of scatter about the calibration line, indicating poor experimental design.

Table 2
Bootstrap results for nickel parameters

Parameter	True value	95% BS confidence interval
α	670	(653, 689)
β	10.3	(10.2, 11.1)
σ_ϵ	43.3	(23.1, 61.1)
σ_η	0.226	(0.133, 0.314)
T_{gf}	1.19	(−0.600, 0.844)
S_{gf}	1.96	(−0.0835, 0.118)

bootstrap confidence interval, showing that poor experimental design can result in lack-of-fit; see Table 2. This particular experimental design results in underestimation of the variance at each concentration. The experimental design statistic, S_{gf}, lies well above the bootstrap confidence interval, confirming that the lack-of-fit is probably due to poor experimental design. Here again both α and β are well estimated by the MLE routine. The MLE routine produces reasonable estimates of the variance components in spite of poor experimental design, see Table 2.

7.4. Conclusion

The two-component error model is useful for many applications in the assessment of environmental data since it provides accurate estimates of error across the entire usable range of a measurement technology, so long as the data exhibit the error structure specified by the model. The model has been tested on a wide variety of data sets, two of which were shown here. The estimation routine produces highly accurate maximum likelihood estimates for the model parameters for each of the data sets tested.

The two-component error model is especially useful in the calculation of critical values and limits of detection based on standard probability theory and also allows calculation of limits of quantification. Thus, the model provides a solid analytical framework for making detection decisions and a superior alternative to previous methods for calculating the quantities mentioned above. That is, the model facilitates explicit evaluation of the efficacy of alternative values for RSD as a criterion of quantification.

The two-component model provides reasonable and reliable method for obtaining parameters estimates for calibration curve data that have desirable statistical properties using well-understood principles of probability. The MLE algorithm produces excellent results as seen in simulations. Goodness-of-fit statistics which test for overall fit of the data and valid experimental design can be useful in identifying data which do not fit the model or were not properly randomized, respectively. Estimated distributions for the parameter estimates, along with confidence intervals for use in the goodness-of-fit tests can be constructed using the parametric bootstrap.

Acknowledgements

Research reported in this paper was supported by grants from the United States Environmental Protection Agency (CR 825621-01-0) and the National Institute of Environmental Health Sciences (P42 ES 04699). The data used as examples are courtesy of the National Council of the Paper Industry for Air and Stream Improvement (organics) and the US EPA (metals), and partially by Financial Assistance Award Number DE-FC09-96SR18546 from the DOE to the University of Georgia Research Foundation. The authors wish to thank the two anonymous referees for their critical comments and constructive suggestions which improved the paper considerably. This work was supported in part by contract DE-FC09-96SR18546 between the US Department of Energy and the University of Georgia's Savannah River Ecology Laboratory.

Appendix A. Closed form solution for IUPAC minimum detectable value

Given a critical level L_C with associated confidence level δ, and a desired confidence level δ' for L_D, the IUPAC minimum detectable value is the true concentration L_D such that the measured response will exceed L_C at least a fraction $1-\delta'$ of the time. The exact solution of this problem under the two-component model can be difficult, but a good approximation can be derived using normal theory by letting L_D be the solution to the following equation:

$$(\alpha + \beta L_D) - z_1\sqrt{\mathrm{Var}(y; L_D)} = L_C = \alpha + z_0\sigma_\epsilon \tag{A.1}$$

$$\beta L_D - z_1\sqrt{\mathrm{Var}(y; L_D)} = z_0\sigma_\epsilon \tag{A.2}$$

$$L_D - z_1\sqrt{V(\hat{\mu}; L_D)} = z_0 S_\epsilon \tag{A.3}$$

where $z_0 = z_\delta$ and $z_1 = z_{\delta'}$, the appropriate quantiles of the normal distribution:

$$L_D - z_1\sqrt{L_D^2 S_\eta^2 + S_\epsilon^2} = z_0 S_\epsilon \tag{A.4}$$

has a solution for L_D exactly when

$$S_\eta < \frac{1}{z_1} \tag{A.5}$$

The solution is given by

$$L_D = \frac{S_\epsilon\left[z_0 + \sqrt{z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)}\right]}{1 - z_1^2 S_\eta^2}. \tag{A.6}$$

If $z_0 = z_1 = z$, the solution has the particularly simple form

$$L_D = \frac{2zS_\epsilon}{1 - z^2 S_\eta^2}. \tag{A.7}$$

A.1. Proof

It is almost immediately apparent that L_D need not exist under the two-component model, or any model that allows the error variance to increase with the concentration. If $z_1\sqrt{V(\mu)}$ exceeds μ for every value of μ, then Eq. (A.3) certainly cannot be solved. This in turn will be true if

$$z_1^2\, e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) > 1 \tag{A.8}$$

$$z_1^2 S_\eta^2 > 1 \tag{A.9}$$

$$S_\eta > \frac{1}{z_1} \tag{A.10}$$

Thus, it is necessary for L_D to exist that $S_\eta < 1/z_1$, and it also turns out to be sufficient.

We can solve the defining equation for L_D in the following way:

$$-z_1\sqrt{L_D^2 S_\eta^2 + S_\epsilon^2} = z_0 S_\epsilon - L_D \tag{A.11}$$

$$z_1^2 L_D^2 S_\eta^2 + z_1^2 S_\epsilon^2 = (z_0 S_\epsilon - L_D)^2 \tag{A.12}$$

$$z_1^2 L_D^2 S_\eta^2 + z_1^2 S_\epsilon^2 = z_0^2 S_\epsilon^2 - 2 z_0 S_\epsilon L_D + L_D^2 \tag{A.13}$$

or

$$L_D^2(1 - z_1^2 S_\eta^2) - L_D(2 z_0 S_\epsilon) + (z_0^2 - z_1^2) S_\epsilon^2 = 0. \tag{A.14}$$

It is not difficult to show that this quadratic equation will have real solutions whenever (A.5) holds. The discriminant of (A.14) is

$$D = 4 z_0^2 S_\epsilon^2 - 4(1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2) S_\epsilon^2 \tag{A.15}$$

$$D = 4 S_\epsilon^2 [z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)]. \tag{A.16}$$

This can only be negative if the second term within square brackets is negative (since the first term is positive). By hypothesis, $1 - z_1^2 S_\eta^2 > 0$, so that the only way this term can be negative is when $z_0^2 > z_1^2$. However, $0 < 1 - z_1^2 S_\eta^2 < 1$, so under these conditions,

$$z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2) > z_0^2 - (z_0^2 - z_1^2) = z_1^2 > 0 \tag{A.17}$$

Hence, the discriminant is always positive when condition (A.5) is satisfied.

The appropriate solution to this quadratic equation satisfies

$$L_D = \frac{(2 z_0 S_\epsilon) + 2 S_\epsilon \sqrt{z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)}}{2(1 - z_1^2 S_\eta^2)} \tag{A.18}$$

$$L_D = \frac{S_\epsilon \left[z_0 + \sqrt{z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)} \right]}{1 - z_1^2 S_\eta^2} \tag{A.19}$$

and this solution is positive exactly under our hypothesis. If $z_0 = z_1$, the solution has the particularly simple form

$$L_D = S_\epsilon \frac{2 z_0}{1 - z_1^2 S_\eta^2} \tag{A.20}$$

Intuitively, the limit of detection would be $S_\epsilon(z_0 + z_1)$ if the variance were constant, and the factor in the denominator inflates the result to account for the increasing variance.

Let us consider an example. Suppose that $\alpha = 0$ and $\beta = 1$, so that the concentration and the response are on the same scale, and suppose that the measurement process parameters are $\sigma_\epsilon = S_\epsilon = 1$ and $\sigma_\eta = 0.1$. Then

$$S_\eta = \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)} = 0.10075, \tag{A.21}$$

which is essentially the same as σ_η. When the confidence level for both the critical value and the detection limit are chosen to be 95%, that is $z_0 = z_1 = 1.645$, application of the formulas leads to $L_C = 1.645$ and $L_D = 3.383$ (only slightly greater than if the variance did not increase with the concentration). If a confidence level of 99% for both is desired then, $z_0 = z_1 = 2.326$ and $L_C = 2.326$ and $L_D = 4.923$.

On the other hand, if $\sigma_\eta = 0.3$, then $L_C = 2.326$ as before, but L_D now increases to 10.518. Finally, when $\sigma_\eta = 0.385$, we cannot find a solution for L_D since $S_\eta = 0.4305 > 1/2.326 = 0.4299$. When the variance increases this rapidly with the concentration, no concentration can guarantee at the 99% confidence level that the measured value will exceed L_C.

References

[1] L. Currie, Limits for qualitative detection and quantitative determination—application to radiochemistry, Anal. Chem. 40 (3) (1968) 586.

[2] A. Hubaux, G. Vos, Decision and detection for linear calibration curves, Anal. Chem. 42 (8) (1970) 849–855.

[3] D.M. Rocke, S. Lorenzato, A two-component model for measurement error in analytical chemistry, Technometrics 37 (2) (1995) 176–184.

[4] M.E. Zorn, R.D. Gibbons, W.C. Sonzogni, Weighted least-squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration, Anal. Chem. 69 (1997) 3069–3075.

[5] M.E. Zorn, R.D. Gibbons, W.C. Sonzogni, Evaluation of approximate methods for calculating the limit of detection and the limit of quantification, Environ. Sci. Technol. 33 (1999) 2291–2295.

[6] R.D. Gibbons, Statistical Methods for Groundwater Monitoring, Wiley, New York, 1994.

[7] A.H. Grange, C.B. William, Determining elemental compositions from exact masses and relative abundances of ions, Trends Anal. Chem. TrAC-15 (1) (1996) 12–17.

[8] C. Liteanu, I. Rica, Statistical Theory and Methodology of Trace Analysis, Ellis Horwood, Chichester, UK, 1980.

[9] L. Oppenheimer, T.P. Capizzi, R.M. Weppelman, H. Mehta, Determining the lowest limit of reliable assay measurement, Anal. Chem. 55 (4) (1983) 638–643.

[10] L. Currie, Nomenclature in evaluation of analytical methods including detection and quantification capabilities, Pure Appl. Chem. 67 (10) (1995) 1699–1723.

[11] R.O. Gilbert, Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, New York, 1987.
[12] A.C. Cohen, Truncated and Censored Samples, Marcel Dekker, New York, 1991.
[13] S. Kuttatharmmakul, J. Smeyers-Verbeke, D.L. Massart, D. Coomans, S. Noack, The mean and standard deviation of data, some of which are below the detection limit: an introduction to maximum likelihood estimation, Trends Anal. Chem. TrAC-19 (4) (2000) 215–222.
[14] L.A. Currie, Detection: international update, and some emerging dilemmas involving calibration, the blank, and multiple detection decisions, Chemomet. Intell. Lab. Syst. 37 (1) (1997) 151–181.
[15] P.C. Meier, R.E. Zünd, Statistical Methods in Analytical Chemistry, Wiley, New York, 1993.
[16] R.D. Gibbons, D.E. Coleman, R.F. Maddalone, An alternative minimum level definition for analytical quantification, Environ. Sci. Technol. 31 (1997) 2071–2077.
[17] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.