Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecotoxicology and Environmental Safety 56 (2003) 78–92
Ecotoxicology and Environmental Safety
http://www.elsevier.com/locate/ecoenv

Modeling uncertainty in the measurement of low-level analytes in environmental analysis

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

[a] *Department of Applied Science and Division of Biostatistics, University of California, Davis, CA 95616, USA*
[b] *US Environmental Protection Agency, USA*

Received 19 March 2003; accepted 19 March 2003

Abstract

The use of analytical chemistry measurements in environmental monitoring is dependent on an assessment of measurement error. Models for variation in measurements are needed to quantify uncertainty in measurements, set limits of detection, and preprocess data for more sophisticated analysis in prediction, classification, and clustering. This article explains how a two-component error model can be used to accomplish all of these objectives. In addition, we present applications to quantitating biomarkers of exposure to toxic substances using gene expression microarrays.

1. Introduction

Limitations of the analytical methodology used to measure the concentration of toxic substances in the environment have had an important role in environmental modeling and regulation. It is difficult to model or 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 policymaking. Furthermore, even when the measurements are reliably above zero, both modeling and regulation are difficult without "error bars" around the measurements. In this article, we discuss the implications of a model for measurement error for these 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 approximately constant. Over ranges of higher concentration, the measurement error is observed to be proportional to the concentration of the analyte (Currie, 1968; Hubaux and Vos, 1970). 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. Because environmental monitoring data often fall into this gray area, understanding measurement error in this region is of considerable importance.

The model presented here resolves these difficulties 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—one of zinc by inductively coupled plasma mass spectrometry (ICPMS) and one of propionitrile by gas chromatography–mass spectrometry (GC-MS). These examples support the validity and 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. Some of the technical background to this model can be found in Rocke and Lorenzato (1995).

*Corresponding author. Fax: +1-530-752-8894.
E-mail address: dmrocke@ucdavis.edu (D.M. Rocke).

0147-6513/03/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0147-6513(03)00052-6

We also present here some recent work on the application of this model to biomarkers of exposure to toxic substances using gene expression data. We find that the perspective of the two-component model clarifies several of vexing problems with gene expression data and allows for much more precise quantification of such biomarkers (Rocke and Durbin, 2001; Durbin et al., 2002).

2. The model

Most measurement technologies require a linear calibration curve 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

$$y = \alpha + \beta\mu e^{\eta} + \varepsilon, \tag{2.1}$$

where y is the response (such as peak area) at concentration μ, $\eta \sim N(0, \sigma_\eta)$, and $\varepsilon \sim N(0, \sigma_\varepsilon)$. Here, η represents the proportional error that always exists, but is noticeable at concentrations significantly above zero, and ε represents the additive error that always exists but is noticeable 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. To obtain the estimated concentration, we do the back calculation:

$$\hat{\mu} = \frac{y - \hat{\alpha}}{\hat{\beta}}, \tag{2.2}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β, respectively, in model (2.1).

Under this model, the variance of the response y at concentration μ is given by

$$\mathrm{Var}\{y\} = \mu^2\beta^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\varepsilon^2 \tag{2.3}$$

(Rocke and Lorenzato, 1995). Two derived quantities will be useful in interpretation of the results. $S_\varepsilon = \sigma_\varepsilon/\beta$ represents the approximate standard deviation of $\hat{\mu}$ at low levels (approximate because it ignores uncertainty in the calibration parameters α and β). $S_\eta = \sqrt{e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)}$ is the approximate RSD of $\hat{\mu}$ for high levels. For values of σ_η appropriate for analytical technologies (say no 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

$$\mathrm{Var}\{y\} = \mu^2\beta^2 S_\eta^2 + \sigma_\varepsilon^2 \tag{2.4}$$

and for the estimated concentration,

$$\mathrm{Var}\{\hat{\mu}\} = \mu^2 S_\eta^2 + S_\varepsilon^2. \tag{2.5}$$

Let us compare the usefulness of a two-component error model to a model using only the relative standard deviation (RSD), which is defined to be equal to the standard deviation of the estimated concentration divided by the concentration (Liteanu and Rica, 1980). For the two-component model, we have

$$\mathrm{RSD}\{\hat{\mu}\} = \sqrt{S_\eta^2 + S_\varepsilon^2/\mu^2}. \tag{2.6}$$

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. 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, critical levels, and quantification levels.

As an example (zinc by ICPMS described in detail later), suppose that $\sigma_\varepsilon = 204$ (in units of the peak area), $\sigma_\eta = 0.0390$ (this is the high-level CV), $\alpha = 490$, and $\beta = 7.06$. The derived quantities are $S_\varepsilon = 204/7.06 = 28.9$ ppt and $S_\eta = 0.0390$. Then the standard deviation of blanks is 204 (units of peak area) or 28.9 ppt. Critical levels, which are the basis for determining detection of an analyte, are often set at 2–3 times the standard deviation of the blank above background (see Currie, 1995, 1997). Using this definition with a multiplier of 3, we have the critical level set at $490 + 3(204) = 1102$ in units of peak area or $3(28.9\text{ ppt}) = 86.7$ ppt. Then, using (2.4), the standard deviation of the response, y, at concentration $\mu = 86.7$ ppt is

$$\begin{aligned}\mathrm{SD}\{y\} &= \sqrt{(86.7)^2(7.06)^2(0.0390)^2 + (204)^2} \\ &= \sqrt{42,200} \\ &= 205\end{aligned}$$

(compared to 204 at zero concentration). The standard deviation of the estimated concentration is, using (2.5),

$$\begin{aligned}\mathrm{SD}\{\hat{\mu}\} &= \sqrt{(86.7)^2(0.0390)^2 + (28.9)^2} \\ &= \sqrt{847} \\ &= 29.1\text{ ppt}\end{aligned}$$

(compared to 28.9 ppt at zero concentration).

That is, using the two-component model, measurements at this concentration have a standard deviation of $\sim$29.1 ppt, only slightly above the value for blanks, and $\mathrm{RSD} = 29.1/86.7 = 0.34$.

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 are replicated. The parameters σ_η can be likewise estimated from the standard deviation of the logarithms of high-level measurements. The calibration curve can then be estimated using weighted least squares, weighting each point by the inverse estimated variance. 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 nonconvergence or impossible estimates (such as negative variances).

The most effective estimation method is maximum likelihood, as described in Rocke and Lorenzato (1995). A computer program that solves for the maximum likelihood estimates for $\alpha, \beta, \sigma_\varepsilon, \sigma_\eta$ is available at `http://www.cipic.ucdavis.edu/~dmrocke`.

4. Applications

In this section, we describe some applications of the two-component model. Special emphasis is given to applications in environmental monitoring, where detection and measurement of toxins at very low levels can be quite important.

4.1. Critical levels

Detection refers to the capacity 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. 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, whereas a result lower than this value indicates that the true concentration in the sample either is 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 (Currie, 1995). 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. Because 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 article.

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 $L_C = \alpha + z_0\sigma_\varepsilon$, 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, that is,

$$L_C = \alpha + z_0\sigma_\varepsilon \tag{4.1}$$

in units of the response and

$$L_C = z_0 S_\varepsilon \tag{4.2}$$

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.

For our example zinc data, we have $\sigma_\varepsilon = 204$ in units of peak area and $S_\varepsilon = 28.9$ ppt. 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 $(2.326)(28.9) = 67.2$ ppt.

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 (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 value from the loss of information; this 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 sophisticated methods can be used to cope with these censored data (Gilbert, 1987; Cohen, 1991), 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. 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 (2.4) and solve the resulting equation. Recall that, from (2.5),

$$\mathrm{Var}\{\hat{\mu}\} = \mu^2 S_\eta^2 + S_\varepsilon^2 \tag{4.3}$$

so that L_D is the solution to the equation

$$L_D = z_0 S_\varepsilon^2 + z_1\sqrt{\mathrm{Var}\{L_D\}} \tag{4.4}$$

$$= z_0 S_\varepsilon^2 + z_1\sqrt{L_D^2 S_\eta^2 + S_\varepsilon^2}, \tag{4.5}$$

where z_0 is the percentile of the standard normal distribution corresponding to the desired confidence level for the critical level, and z_1 is the percentile corresponding to the desired confidence level for the minimum detectable value. Because L_D appears on both sides of the equation, at first glance it might seem that iterative methods would be required to find the solution. However, so long as the variance of y does not increase too rapidly with μ, a closed form solution is possible. The required condition is

$$S_\eta < 1/z_1. \tag{4.6}$$

For details and proof see Wilson et al. (2001). The solution derived there is

$$L_D = \frac{S_\varepsilon[z_0 + \sqrt{z_0^2 - (1 - z_1^2 S_\eta^2)(z_0^2 - z_1^2)}]}{(1 - z_1^2 S_\eta^2)}. \tag{4.7}$$

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

$$L_D = \frac{2zS_\varepsilon}{1 - z^2 S_\eta^2}. \tag{4.8}$$

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 $S_\varepsilon = 28.9$ 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

$$(2)(2.326)(28.9)/[1 - (2.326)^2(0.0390)^2] = 135 \text{ ppt}. \tag{4.9}$$

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. Because the standard deviation at low levels is actually smaller than that at high levels, it could be argued that this is a meaningless concept; any measured value, along with a confidence interval, provides a useful measurement; and 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 (RSD) falls to a specified level (Gibbons, 1994; Meier and Zünd, 1993). 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 factor of

σ_η, say $2\sigma_\eta$. It makes little sense to specify a particular arbitrary RSD, such as 20%, because 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, for example, 10%. Then the equation

$$\sqrt{V(L_Q)/L_Q} = R, \tag{4.10}$$

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

$$L_Q = \sqrt{\frac{S_\varepsilon^2}{R^2 - S_\eta^2}} \tag{4.11}$$

whenever $R > S_\eta$. No real solution is possible when $R \leqslant S_\eta$. This is readily apparent once (4.10) is rewritten as a quadratic equation in L_Q.

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

$$L_Q = 28.9/\sqrt{(0.1)^2 - (0.0390)^2} = 314 \text{ ppt}. \tag{4.12}$$

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. Whereas 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.

4.4. Sample size calculations

When detection is crucial, it makes sense to take replicate samples at a site rather than one measurement. The measurement error associated with the mean of the replicates is much smaller than the error associated with one measurement, thus replication allows for more sensitive detection of the toxin. Therefore, sample size calculations become quite important.

According to our model, the measurements at true concentration $\mu = 0$ are normally distributed with standard deviation σ_ε. If the number of replicates is r, then any average of measured values greater than

$$D = \alpha + \frac{3\sigma_\varepsilon}{\sqrt{r}} \tag{4.13}$$

is extremely unlikely to have come from a zero concentration sample. Similarly, so is any average of estimated concentrations greater than

$$D = \frac{3S_\varepsilon}{\sqrt{r}}. \tag{4.14}$$

(Of course, this assumes that the replicates are true reruns of the entire process; otherwise the error may not be reduced by a factor of $\sqrt{r}$, but by a much smaller amount.)

An implication of this rule for environmental monitoring is that multiple measurements at low levels, even if they are individually below the critical level, can still provide quantitative evidence of the concentration of a toxic substance. If the safe level is near or below the critical level, then requiring adequate replication of the measurements reduces the effective critical level. That is, much smaller amounts can be measured with greater certainty. Quantitative recording of repeated measurements, even when they are below the individual observation critical level, will allow quantitative evidence to be gathered to estimate the true concentration at a site.

The question remains, how many replicates are needed? The approximate sample size required to determine a concentration to a particular precision depends on the concentration as well as the precision desired. Suppose that we wish to determine whether the concentration of zinc is above some specific level, say 50 ppt. From previous experience, we estimate the standard deviation parameters to be $S_\varepsilon = 28.9$ ppt and $S_\eta = 0.0390$. Suppose further that it is held to be important to detect a concentration of 80 ppt and determine it to be out of compliance. At that concentration, the standard deviation of an average of r replicates is

$$\begin{aligned}\sqrt{\frac{S_\varepsilon^2 + \mu^2 S_\eta^2}{r}} &= \sqrt{\frac{(28.9)^2 + (80)^2(0.0390)^2}{r}} \\ &= \sqrt{\frac{845}{r}} \\ &= \frac{29.1}{\sqrt{r}}.\end{aligned} \tag{4.15}$$

Using a normal approximation, the distance between the criterion level 50 ppt and the concentration 80 ppt in standard deviation units is

$$(30/29.1)\sqrt{r} = 1.03\sqrt{r}. \tag{4.16}$$

For the chance of detection to be 0.95, we require

$$1.03\sqrt{r} > 1.645$$

or $r > 2.55$. In this case, then, we need at least three replicates.

4.5. Uncertainty of a single measurement

The uncertainty of a single measurement is usually quantified using confidence intervals. There are two primary approaches to this problem, an exact solution and a normal or log normal approximation. The exact solution requires numerical integration and will not be discussed here. Say we would like a 95% confidence interval for μ based on a single measurement, $\hat{\mu}$. The approximate method for low values of $\hat{\mu}$, using an estimated variance and a normal approximation is

$$\hat{\mu} \pm 1.96\sqrt{\mathrm{Var}(\hat{\mu})}, \tag{4.17}$$

where $\mathrm{Var}(\hat{\mu})$ is estimated using

$$\mathrm{Var}(\hat{\mu}) = \mu^2 S_\eta^2 + S_\varepsilon^2 \tag{4.18}$$

and where all estimates are obtained from the maximum-likelihood routines or other methods. For high levels of $\hat{\mu}$ (those in which the second term in Eq. (2.5) dominates), $\ln \hat{\mu}$ is approximately normally distributed with variance σ_η^2. Hence, a 95% confidence interval for μ is

$$(\exp(\ln \hat{\mu} - 1.96\hat{\sigma}_\eta), \exp(\ln \hat{\mu} + 1.96\hat{\sigma}_\eta)). \tag{4.19}$$

Note that this interval is symmetric on the log scale, but asymmetric on the original measurement scale.

For our zinc data, a measured concentration of 80.0 ppt would use the first technique and have a 95% confidence interval of $80.0 \pm (1.96)(29.1) = 80 \pm 57.0$ (because the standard deviation of the 80.0-ppt measured value is $\sqrt{(28.9)^2 + (80)^2(0.0390)^2} = 29.1$). A measured value of 5000 ppt would use the second method. The log measurement is 8.517 with standard deviation $\sigma_\varepsilon = 0.0390$, so a 95% confidence interval on the log scale is $8.517 \pm (1.96)(0.0390) = 8.517 \pm 0.076$, or (8.441,8.593). Transforming the limits by exponentiation, we arrive at the 95% confidence limit (4630,5390) for the concentration.

We can also use this method to give confidence intervals for the average of a series of replicate measurements. For low levels, the average of r measurements will be approximately normally distributed with standard deviation $\sqrt{\mathrm{Var}(\hat{\mu})/r}$. For larger values of $\hat{\mu}$, the average of the natural log of the r measurements will have approximate standard deviation $\sigma_\eta/\sqrt{r}$. Confidence intervals can then be constructed as described earlier using the appropriate standard deviations.

An alternative to the use of the raw or log data is a new data transformation that can be used over the whole range of the data. This is described in the next section and documented fully in Hawkins (2002) and Durbin et al. (2002).

5. Data transformations

Transformations of statistical data are undertaken for a variety of reasons. It may be that transformed data conform to a linear or additive model more accurately, they may have more nearly constant variance, or may be more nearly normally distributed. One of the most important such goals is stabilisation of variance, so that all of the observations have the same variability.

If the variance is already constant, as is approximately true for low-level analytical data, then no data transformation is required. If the standard deviation is proportional to the mean (that is, the relative standard deviation is constant), then the log transform stabilizes the variance. In the case of analytical data spanning the whole range, neither the log transformation nor the raw data satisfy the assumption of constant variance.

As it turns out, there is a transformation that accomplishes this goal. Suppose that there are random variables y_i that estimate μ_i, and suppose that $\mathrm{Var}(y_i) = \sigma_0^2 v(\mu_i)$. Consider a transformation $z = f(y)$. It is well known that, up to the first order,

$$\mathrm{Var}(z_i) = (f'(\mu_i))^2 \sigma_0^2 v(\mu_i).$$

Thus, a transformation that fully stabilizes the variance would be one in which

$$(f'(\mu_i))^2 = \frac{1}{\sigma_0^2 v(\mu_i)}$$

or

$$f'(\mu_i) = \frac{1}{\sqrt{\sigma_0^2 v(\mu_i)}}.$$

This formulation gives us many familiar results. If $v(\mu) = \mu^2$, then

$$f'(\mu) = \frac{1}{\mu}$$

so

$$f(\mu) = \ln(\mu).$$

If $v(\mu) = \mu$ (as in Poisson data), then

$$f'(\mu) = \mu^{-1/2},$$

so

$$f(\mu) = \sqrt{\mu}.$$

If

$$v(y) = a^2 + b^2\mu^2,$$

where $a = S_\varepsilon$ and $b = S_\eta$, as is the case with analytical data across the whole range, we have

$$f'(\mu) = \frac{1}{\sqrt{a^2 + b^2\mu^2}},$$

which integrates to a multiple of

$$f(\mu) = \ln(\mu + \sqrt{\mu^2 + a^2/b^2})$$

(Hawkins, 2002; Durbin et al., 2002). We then have that $z = f(y)$ has constant approximate standard deviation of $b = S_\eta$ over the whole range. This is the same standard deviation as $\ln(y)$ has for high-level values, but this transformation achieves the same standard deviation across low levels as well.

We can use this to set a confidence interval that works for any concentration. The inverse transformation of

$$z = f(y) = \ln(y + \sqrt{y^2 + a^2/b^2})$$

is

$$y = g(z) = (e^y - (a^2/b^2)e^{-y})/2.$$

For the zinc data, $a^2/b^2 = 549,119$. If we have a measured value of 1000 ppt, then the transformed value is $f(1000) = \ln(1000 + \sqrt{1000^2 + 549,119}) = 7.716$. A 95% confidence interval is $7.716 \pm (1.960)(0.0390)$, or (7.640,7.692). Transforming back to the original scale with the g inverse transformation gives (908,1098).

For low level data, the new transformation leads to similar results to using the raw data. For $\hat{\mu} = 80$ ppt, the raw-data confidence interval is (23, 137), whereas the transformation yields (23,137), the same up to whole units of ppt. For a large value such as 5000 ppt, the use of the log transform gives a confidence interval of (4632, 5397), whereas the new transformation yields (4628, 5401), the same to three significant figures.

Table 1
Zinc parameter estimates

Parameter	Maximum–likelihood estimate (MLE)
α	490
β	7.06
σ_ε	204
σ_n	0.0390
S_ε	28.9 ppt
S_n	0.0390
L_C (peak area)	965
L_C (concentration)	67.1 ppt
L_D	135 ppt
L_Q	314 ppt

The MLEs were obtained using numerical methods in Fortran and then used to calculate the critical value (L_C) limit of detection (L_D) and limit of quantification (L_Q). The first value reported for L_C is peak area; the second is in units concentration (parts per trillion). The L_D and the L_Q are both in units of concentration. L_Q was calculated using a CV of 0.10. Both L_D and L_C were calculated using a 99% confidence level.

Table 2
Zinc data set

Concentration in parts per thousand	Number of replicates	Average observed response (peak area)	Predicted response
0	8	265	490
10	7	317	561
20	7	692	631
100	11	1290	1200
200	7	2190	1900
500	7	4110	4020
1000	9	7590	7550
2000	7	14,400	14,600
5000	9	35,100	35,800
10,000	10	70,300	71,100
25,000	9	181,000	177,000

The predicted response at each concentration was calculated using the maximum-likelihood estimates for the parameters of the calibration curve.

6. Instrumental method examples

In this section, we present two examples of analytical methods and examine the fit of the two-component error model. Our first example is a metal data set using ICPMS. Our second example is an organic data set using GC-MS. In both cases, we use the calibration curve model (2.1).

6.1. Zinc by ICPMS

This first example is a set of zinc data at 12 concentrations measured via IPCMS, with 7–11 replications at each concentration. The maximum-likelihood parameter estimates for α, β, σ_ε and σ_η are shown in Table 1. The summary data are shown in Table 2. The observed and predicted responses are shown in Fig. 1. A plot of the residuals vs. predicted values is shown in Fig. 2. Note the increasing deviation of the observed response away from the calibration curve (that is, the predicted response). Thus, the data exhibit the error structure assumed by the model. The observed and model predicted variances are shown in Fig. 3. Here also, the linear increase in standard deviation at high levels is clear, as well as an area at concentrations near zero where the standard deviation appears to remain somewhat constant.

6.2. Propionitrile by GS-MS

Our second example is a set of propionitrile data measured at nine concentrations, using GC-MS, with the number of replications varying between 4 and 11. The maximum-likelihood estimation results are shown in Table 3. The summary data are shown in Table 4. The observed and predicted responses are shown in Fig. 4.

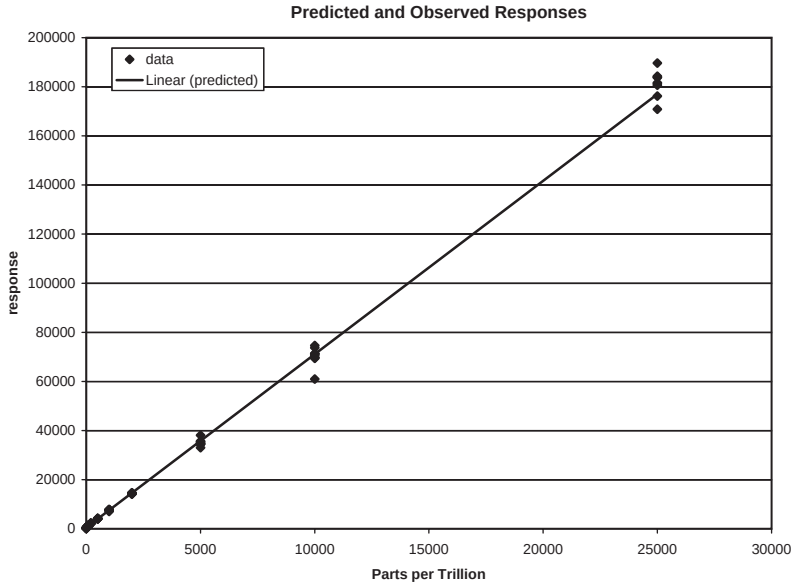

Fig. 1. Zinc—predicted and observed responses.

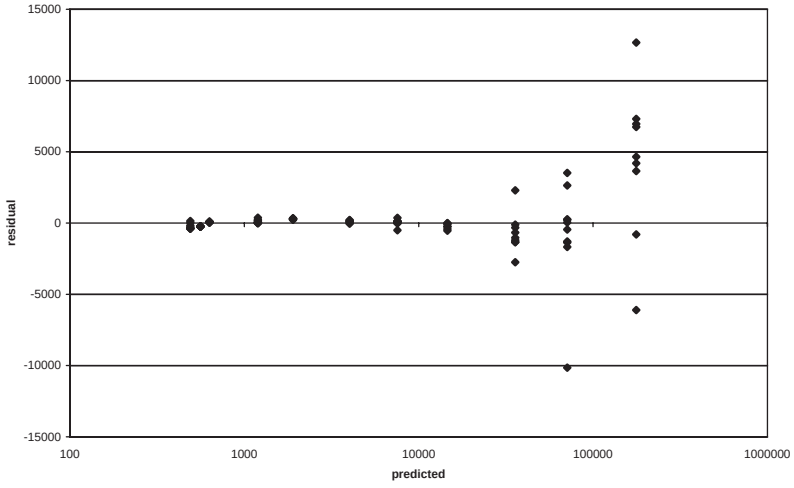

Fig. 2. Zinc residuals.

The residual vs. predicted plot is shown as Fig. 5. The observed and predicted variances are shown in Fig. 6. Once again, it can be easily observed that the data exhibit the error structure of the model and the estimation routine produces plausible results.

7. Applications to biomarkers of exposure to toxic substances via gene expression arrays

Just as with any other analytical technology, measurement of gene expression with cDNA or oligonucleotide arrays have measurement errors. It is commonly observed (e.g., Chen et al., 1997) that the standard deviation of measurements rises in proportion to the expression level. However, this proportionality cannot continue down to genes that are entirely unexpressed because that would imply zero measurement error, which is not observed. The model described in this section was originally developed in the context of instrumental methods of analytical chemistry, but these methods also exhibit the same kind of behavior referenced above (Rocke and Lorenzato, 1995). This model resolves the difficulties 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 also discuss the application of the model to detection limits, categorization of genes as expressed or unexpressed, comparison

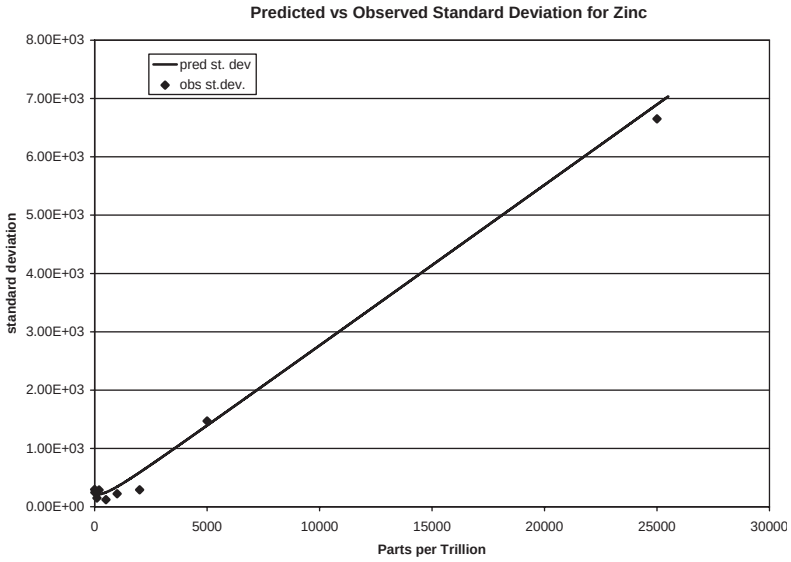

Fig. 3. Predicted vs. observed standard deviation for zinc.

Table 3
Propionitrile parameter estimates

Parameter	Maximum–likelihood estimates
α	559
β	18.7
σ_ε	147
σ_η	0.0397
S_ε	7.85 ppb
S_η	0.0398
L_C (peak area)	900
L_C (concentration)	18.3 ppb
L_D	36.8 ppb
L_Q	85.6 ppb

The MLEs were obtained using numerical methods in Fortran and then used to calculate the critical value (L_C) limit of detection (L_D) and limit of quantification (L_Q). The first value reported for L_C is peak area. The second is in units concentration (parts per billion). The L_D and the L_Q are both in units of concentration. L_Q was calculated using a CV of 0.10. Both L_D and L_C were calculated using a 99% confidence level.

Table 4
Propionitrile data set

Concentration in parts per billion	Number of replicates	Average observed response (peak area)	Predicted response
4.8	7	7.16	7.37
16.0	8	18.1	17.4
32.1	7	31.7	31.7
48.1	3	42.0	46.0
160	6	151	146
321	4	284	289
481	3	431	432
722	5	658	647
3010	4	2660	2680

The predicted response at each concentration was calculated using the maximum-likelihood estimates for the parameters of the calibration curve.

of expression between conditions on the same gene, construction of confidence intervals, and transformation and weighting of expression data for use in comparisons and in multivariate applications such as classification or clustering. Some of the details for these methods are explicated more fully in Rocke and Durbin (2001) and Durbin et al. (2002).

7.1. An error model for gene expression arrays

For gene expression arrays, it is unusual to have calibration data (that is, samples of known expression levels); thus, we cannot actually discern the expression level in molecular units, but can only do so relatively. Model (2.1) then looks like this:

$$y = \alpha + \mu e^{\eta} = \varepsilon, \tag{7.1}$$

where y is the intensity measurement, μ is the expression level in arbitrary units, and α is the mean background (mean intensity of unexpressed genes).[1] Our best estimate if μ is $y - \hat{\alpha}$, the background-corrected observed intensity. The first error term is $\varepsilon \sim N(0, \sigma_\varepsilon)$, which represents the standard deviation of the background (unexpressed genes), and the second error term is $\eta \sim N(0, \sigma_\eta)$, which represents the proportional error

[1] Background is used here for the statistical distribution of overall intensity measurements for genes that are actually not expressed in the sample. We do not discuss here the image processing issued in which background may refer to the pixel distribution in areas of the slide in which there is not spot.

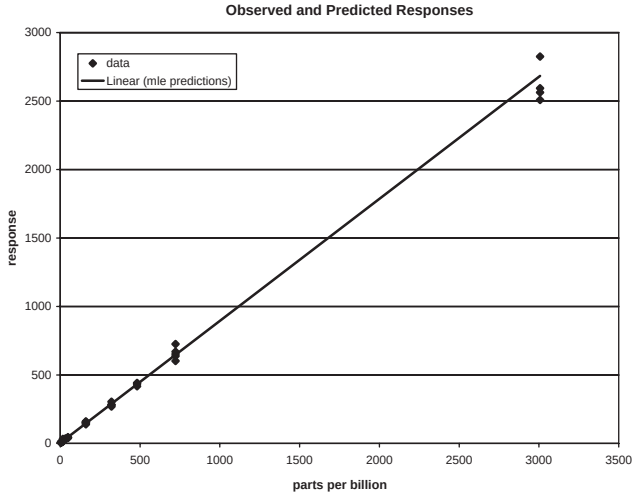

Fig. 4. Propionitrile—observed and predicted responses.

that always exists, but is noticeable mainly for highly expressed genes.

Under this model, the variance of the response y at concentration μ is given by

$$\mathrm{Var}(y) = \mu^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\varepsilon^2 = \mu^2 S_\eta^2 + \sigma_\varepsilon^2. \tag{7.2}$$

We illustrate this with a data set from an experiment, on the response of male Swiss Webster mice to a toxic substance (Bartosiewicz et al., 2000). The treated animal received an intraperitoneal injection of 15 mg/kg of β-naphthoflavone, whereas the control mouse had an injection of the carrier (corn oil) of equal volume. The two-color spotted cDNA slides were constructed using Molecular Dynamics equipment, with data from a treated mouse and a control mouse on each slide. Data were replicated usually a total of eight times

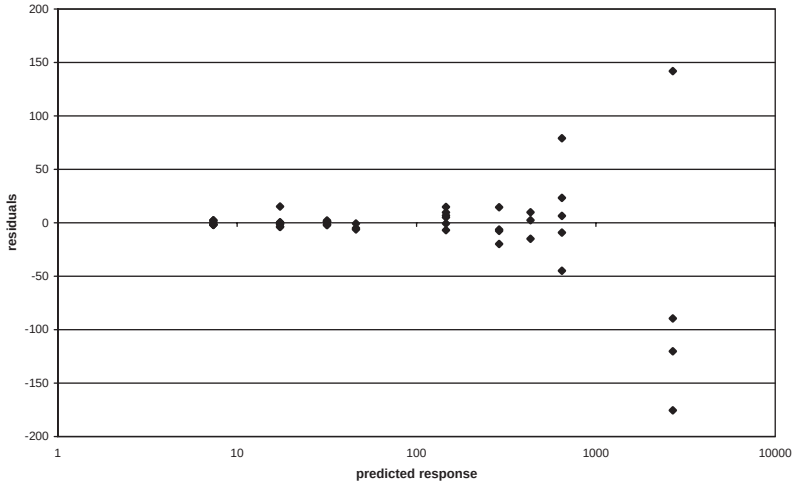

Fig. 5. Propionitrile residuals.

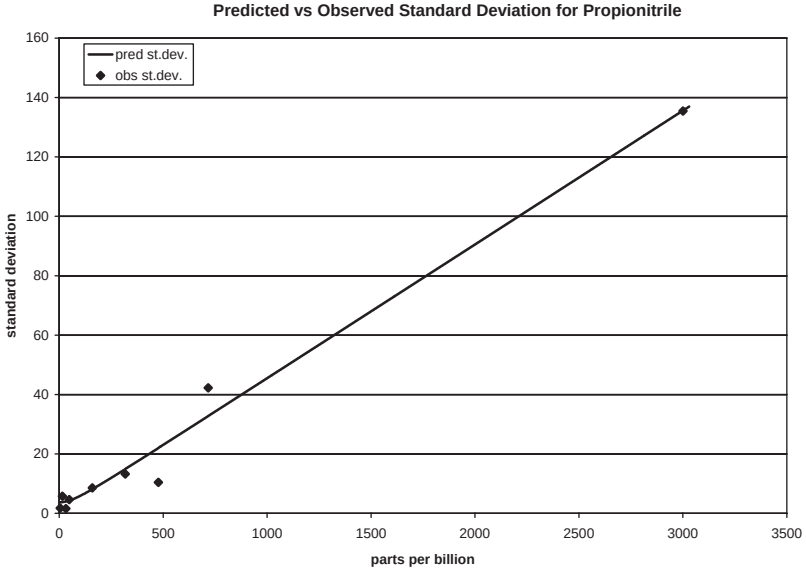

Fig. 6. Predicted vs. observed standard deviation fro propionitrile.

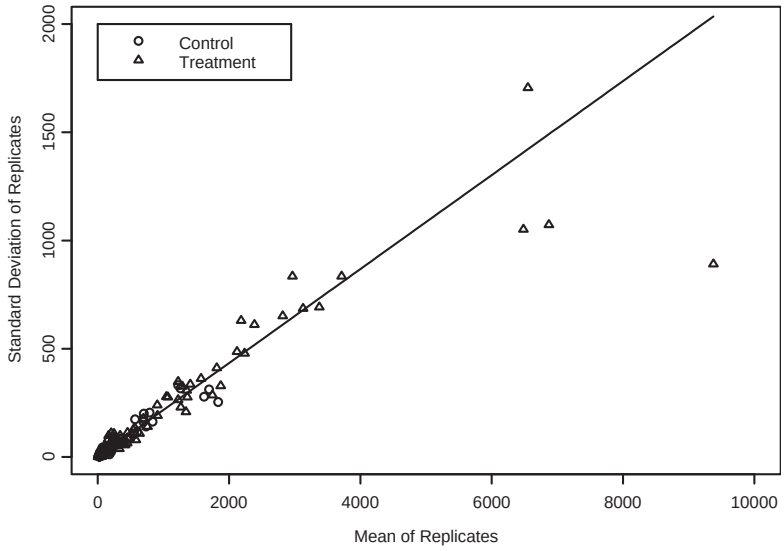

Fig. 7. Raw data (thousands).

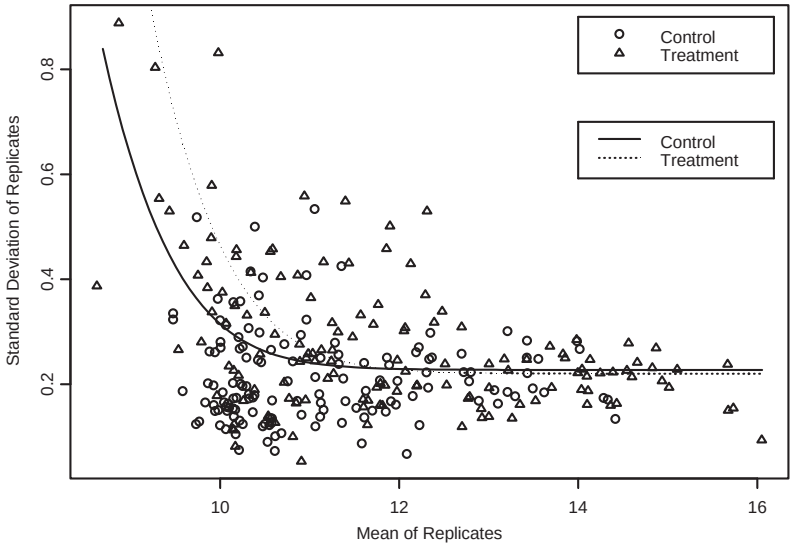

Fig. 8. Logarithms of the data.

per slide (meaning eight spots from the same microplate well were placed on the slide). We use data from one slide. Figs. 7–9 illustrate the phenomena. Fig. 7 shows the close linear relationship between replicate mean and replicate standard deviation at high expression levels. The line shown is the predicted standard deviation from the model; the slope of this line at high levels is the relative standard deviation. Fig. 8 shows the approximately constant standard deviation of the natural logarithms of the data above a log intensity of ~ 13. Note that when the RSD of the original set of replicates is not too large, the standard deviation of the natural logarithms is about the same as the RSD of the untransformed data. This figure also illustrates the vexing phenomenon that the logarithmic transformation, which nicely stabilizes the variance at high levels, produces highly variable results for low expression levels. The lines show the predicted standard deviation from the model for control data (solid line) and treatment data (dotted line) using the model along with the theoretical standard deviation. Fig. 9 shows the approximately constant variance of the raw data below a measurement threshold as predicted by the two-component model.

It is important to realize that most of the variation observed on a cDNA or oligonucleotide array is caused by variations in μ, the actual expression. Variation within replicated spots at the same level μ of true expression is the measurement error that we model, and this is typically much smaller. For our example, the observed mean intensity varies across genes from ~ 6000 to $> 9,000,000$ U. Using the two-component model, the uncertainty in a mean of $9,000,000$ U over eight

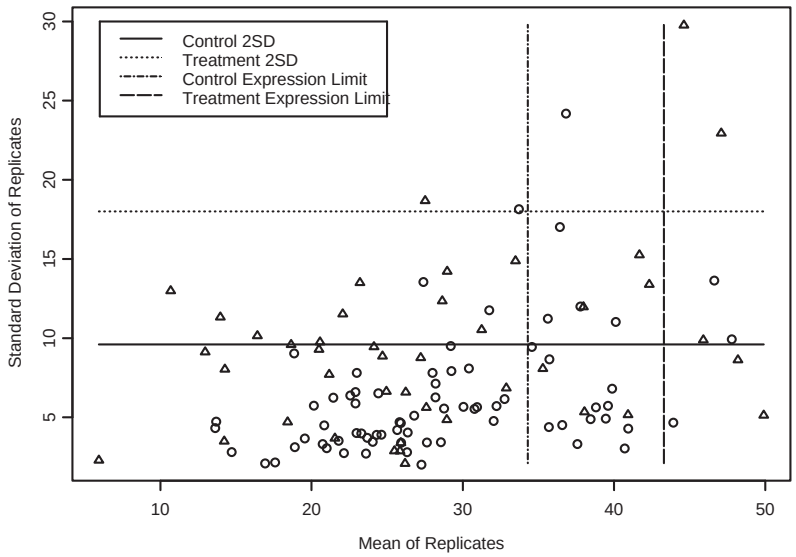

Fig. 9. Raw data for low-level expression (thousands).

replicate amounts to a confidence interval of $\sim 7,650,000$–$10,590,000$ far smaller than the variability between different genes. We also conjecture that variation in μ across individuals or experimental condition for a fixed gene also has a two-component structure in that the standard deviation at low expression levels is approximately constant whereas at high expression levels, the RSD is approximately constant. This issue will be addressed in further work.

7.2. Estimation

A model such as (7.1) cannot be used in practice unless the parameters can be estimated. In this section, we discuss methods of estimation and necessary characteristics of the data for estimation to be possible.

7.2.1. Estimation of background using negative controls

The easiest way to estimate the mean α and standard deviation σ_ε of the low-level measurements is from replicate blanks (negative controls). The standard deviation of the negative controls would be used as the estimate of σ_ε. The mean intensity of the negative controls is a suitable estimate of α, the mean background.

7.2.2. Estimation of background with replicate measurements

If we have replicated measurements, but no specific negative controls, we can still estimate the background mean and variance. According to (7.1), intensity measurements from unexpressed genes will be normally distributed with mean α and standard deviation σ_ε. This can be done separately for treatment and control in a two-color array.

1. Begin with a small subset of genes with low intensity, such as the 10% of genes with lowest intensity measurements. Compute the mean $\bar{x}_B$ of the genes including all replicates and the pooled standard deviation s_B of replicates of these genes that have been replicated. For each replicated gene to this group, compute the standard deviation s_i of the replicates. If there are m replicated genes, pool these estimates as follows:

$$s_B = \sqrt{(n-m)^{-1} \sum_{i=1}^{m} s_i^2 (n_i - 1)},$$

where n_i is the number of replicates for gene i and $n = \sum_{i=1}^{m} n_i$. If there are a large number of such genes, or many replicates, it may be better to use only those replicated genes whose average expression is less than $\bar{x}_B$ to determine s_B.
2. Define a new subset consisting of genes whose intensity values are in the interval $[\bar{x}_B - 2s_B, \bar{x}_B + 2s_B]$. Recompute $\bar{x}_B$ and s_B.
3. Repeat the previous step until the set of genes does not change.

At the final step, the set of genes should include at least 95% of the unexpressed genes. If one includes the genes below $\bar{x}_B - 2s_B$, this would include $\sim 97.5\%$ of the unexpressed genes. Depending on the distribution of actual expression levels, this estimate could be biased up both in the mean and (slightly) in the standard deviation, because it is impossible in principle to distinguish an unexpressed gene from one with such a low expression level that it is below detection limits.

Nonetheless, this estimate should be of considerable use in screening genes for expression.

The standard deviation of an replicated gene is at least σ_ε, and for genes that are unexpressed or expressed at very low levels, it will be essentially exactly this. This process cannot lead to the empty set, because at least one replicated gene will be included. In practice, the process cannot converge to an estimate of α or σ_ε that are systematically too small, and because the bias in the standard deviation at low levels is typically small, the estimates are rarely too large by much. Furthermore, the solution to which this process converges does not appear at all to depend on the details of the selection of the initial set, so long as it is fairly small. These observations have been confirmed by extensive simulations that are omitted here. In the next section, we present a method of estimating α and σ_ε even from unreplicated data.

7.2.3. Estimation of background without replication

In the absence of replicated measurements, it is still possible to estimate the mean and variance of unexpressed genes (background); the following procedure is recommended. This can be done separately for treatment and control in a two-color array.

1. Begin with a small subset of genes with low intensity, such as the 10% of genes with lowest intensity measurements. Compute the mean $\bar{x}_B$ and standard deviation s_B of these genes.
2. Define a new subset consisting of genes whose intensity values are in the interval $[\bar{x}_B - 2s_B, \bar{x}_B + 2s_B]$. Recompute $\bar{x}_B$ and s_B.
3. Repeat the previous step until the set of genes does not change.

At the final step, the set of genes should include at least 95% of the unexpressed genes. As before, this estimate may be biased upward but nonetheless should be of considerable use in screening genes for expression. Note that this procedure is less reliable than the one to be used when there are an adequate number of replicates because the standard deviation is taken across genes. If some of these genes are actually expressed, the standard deviation is elevated by the variation in means, whereas this does not happen with replicated data.

A variant of this procedure may reduce the bias somewhat. In this variant, one uses the median of the expression levels of the subset of genes as the estimate of location, and uses MAD 0.6745 as the estimate of s_B, where the MAD is the median absolute deviation from the median. This is calculated by subtracting the median from each expression value in the subset, taking absolute values, and taking the median of the resultant set of absolute deviations.

7.2.4. Estimation of the high-level RSD

The parameter σ_η can be estimated from the standard deviation of the logarithms of high-level replicated measurements in much the same way as the background standard deviation can be estimated from the low-level data. For each replicated gene that is expressed at a high level, compute the standard deviation s_i of the logarithms of the replicate estimates $\hat{\mu} = y - \hat{\alpha}$ of μ. If there are m replicated genes, one then pools these estimates as follows:

$$s_H = \sqrt{(n-m)^{-1}\sum_{i=1}^{m} s_i^2(n_i - 1)},$$

where n_i is the number of replicates for gene i and $n = \sum_{i=1}^{m} n_i$. This method works because for high-expression levels, (7.1) is indistinguishable from

$$\hat{\mu} = \mu e^{\eta},$$

$$\ln(\hat{\mu}) = \ln(\mu) + \eta,$$

which is a constant mean, constant variance model.

There is no method even in principle for estimating measurement error without at least some replication at high levels because it is impossible from an unreplicated sample to know if an intensity value is high because the expression is high or because of a positive measurement error. This fact of life should be an important determinant of experimental design is microarrays.

7.2.5. What is "high" and "low" expression?

Given the model, and preliminary estimates of the parameters, we can address the issue of high and low expression and the variability of genes in each group. The variance of y given by (7.2) can be compared with the variance of y at low levels. If the ratio is smaller than, say 0.9, then most of the variance is due to the additive error component. Thus,

$$\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \mu^2 S_\eta^2} \geqslant 0.9,$$

$$\sigma_\varepsilon^2 \geqslant 0.9\sigma_\varepsilon^2 + 0.9\mu^2 S_\eta^2,$$

$$\mu^2 \leqslant \frac{0.1\sigma_\varepsilon^2}{0.9 S_\eta^2},$$

$$\mu \leqslant \sigma_\varepsilon / 3S_\eta. \tag{7.3}$$

Thus, one can define "low-level" data as ones for which the observed expression is smaller than this threshold.

Table 5
Parameter estimates for the example data

Parameter	Control	Treatment
$\hat{\alpha}$	24,800	25,300
$\hat{\sigma}_\varepsilon$	4800	9000
$\hat{\sigma}_\eta$	0.227	0.220
$\hat{\sigma}_\eta/\sqrt{r}$	0.080	0.078
S_η	0.236	0.228
Expression cutoff	34,300	43,300

This shows the estimates for the treatment and control data separately; estimates for the combined model are given in Rocke and Durbin (2001). In this case r, the number of replicates, is eight. The expression cutoff is the intensity above which the gene is expressed at a level that is statistically significantly above zero.

Similarly,

$$\frac{\mu^2 S_\eta^2}{\sigma_\varepsilon^2 + \mu^2 S_\eta^2} \geqslant 0.9,$$
$$\mu^2 S_\eta^2 \geqslant 0.9\sigma_\varepsilon^2 + 0.9\mu^2 S_\eta^2,$$
$$\mu^2 \geqslant \frac{0.9\sigma_\varepsilon^2}{0.1 S_\eta^2},$$
$$\mu \geqslant 3\sigma_\varepsilon/S_\eta, \tag{7.4}$$

gives a threshold above which the variance is mostly due to the multiplicative component.

An examination of the example data shows that the variance is approximately constant below $\sim 25,000$ and the variance of the logarithm is approximately constant above ~ 13. Use of the procedure given in Sections 7.2.2 and 7.2.4 yields parameter estimates as given in Table 5.

Using (7.3), for the control data, the logarithms have approximately constant variance when

$$\mu \geqslant 3\sigma_\varepsilon/S_\eta,$$

$$\mu \geqslant 61,000$$

corresponding to a signal of $24,800 + 61,000 = 85,800$ and a log signal of 11.4. For the treatment data, the equivalent values are $\mu \geqslant 118,400$, corresponding to a signal of $25,300 + 118,400 = 143,700$, and a log signal of 11.9.

Using (7.4), the raw control data have approximately constant variance when

$$\mu \leqslant \sigma_\varepsilon/3S_n,$$

$$\mu \leqslant 6800$$

corresponding to a signal of $24,800 + 6800 = 31,600$. For the treatment data, the equivalent calculation gives $\mu \leqslant 13,200$, corresponding to a signal of 38,500. In the range 31,600 to 85,800 for controls and 38,500–143,700 for treatment, both the variance and the coefficient of variation are changing substantially.

For data with calibration curves, the most effective estimation method is maximum likelihood, as described in Rocke and Lorenzato (1995). A method of applying maximum likelihood for replicated microarrays is in development, but the more heuristic methods given here may be satisfactory for many applications.

The recommended course of action at this point is to use the data transformation

$$f(\mu) = \ln(\mu + \sqrt{\mu^2 + a^2/b^2}). \tag{7.5}$$

At that point, standard statistical methods can be used. It is important to note that neither the raw data nor the logarithms can be uniformly used without severe problems. Only this new transformation behaves properly across a wide range of data.

8. Conclusion

The two-component error model is useful for many applications in the assessment of environmental data because 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, three 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 value 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.

Acknowledgments

The United States Environmental Protection Agency (CR 825621-01-0) and the National Institute of Environmental Health Sciences (P42 ES 04699) supported with grants the research on which we report(ed) in this paper. The data we use(d) as examples are courtesy of the National Council of the Paper Industry for Air and Stream Improvement (organics) and the US EPA (metals).

References

Bartosiewicz, M., Trounstine, M., Barker, D., Johnston, R., Buckpitt, A., 2000. Development of a toxicological gene array and quantitative assessment of this technology. Arch. Biochem. Biophys. 376, 66–73.

Chen, Y., Dougherty, E.R., Bittner, M.L., 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics 2 (4), 364–374.

Cohen, A.C., 1991. Truncated and Censored Samples. Marcel Dekker, New York.

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

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

Currie, L.A., 1997. Detection: international update, and some emerging dilemmas involving calibration, the blank and multiple detection decisions. Chemometrics Intell. Lab. Syst. 37 (1), 151–181.

Durbin, B.P., Hardin, J.S., Hawkins, D.M., Rocke, D.M., 2002. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18, S105–S110.

Gibbons, R., 1994. Statistical Methods for Groundwater Monitoring. Wiley, New York.

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

Hawkins, D.M., 2002. Diagnostics for conformity of paired quantitative measurements. Stat. Med. 21, 1913–1935.

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

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

Meier, P.C., Zünd, R.E., 1993. Statistical Methods in Analytical Chemistry. Wiley, New York.

Rocke, D.M., Durbin, B.P., 2001. A model for measurement error for gene expression arrays. J. Comput. Biology 8, 557–569.

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

Wilson, M., Rocke, D.M., Durbin, B.P., Kahn, H., 2001. Application to Environmental Monitoring of a Two-Component Model for Chemical Analytical Error, submitted for publication.