Journal of Data Science **7**(2009), 61-72

Approximate Graphical Methods for Inverse Regression

Geoffrey Jones and Paul Lyons
Massey University

Abstract: Graphical procedures can be useful for illustrating and evaluating the process of inverse regression. We first review some simple and well-known graphical approaches for univariate linear and nonlinear models. We then propose a new graphical tool applicable to situations where the response is bivariate and repeated measures data are available. The proposed method is illustrated with an example of the age determination of tern chicks using measurements on body weight and wing length.

Key words: Calibration, growth curve, longitudinal data, random coefficients.

1. Introduction

In inverse regression, or statistical calibration, the relationship between a response Y and a covariate x is first estimated using a set of training data $(x_1, Y_1), \ldots (x_n, Y_n)$. This relationship is then used to infer the covariate value x_0 corresponding to an observed response Y_0. A comprehensive summary of the theory is given by Brown (1993).

Our interest here is in approximate graphical methods for estimating the unknown x_0 once the relationship between Y and x has been established. It is sometimes the case that a quick approximate answer is all that is required, so that a method which involves merely referring to a graph without the need for computing equipment has much to recommend it, particularly when many estimates are needed in a field situation. Moreover graphical methods can give insights into the nature and reliability of the process, since aspects of statistical uncertainty can be incorporated into the graph.

Our aim in this paper is to illustrate the above points with some fairly well-known approaches, and to introduce a new graphic for inverse regression with bivariate longitudinal data. In the first section we consider inverse regression with a univariate response, focussing on the method of inverting a prediction interval. This is illustrated with some examples chosen to demonstrate the range

of applicability of this method. In the next section the case of bivariate Y is considered. The statistical issues here are more complex, and graphical approaches can provide useful insights. In the third section we consider repeated measures data and introduce a new graphic. We discuss some issues in the construction and use of this graph.

Inverse regression estimation will usually involve adding some new lines to a graph developed from the training data. This may require actually drawing the new lines on the graph, but since we are only expecting approximate answers it may be sufficient to add the new lines "in the mind's eye". These may then be imagined to move, giving a dynamic image of the estimation and perhaps insights into its characteristics. This dynamic use of a graph is a key feature of the approach. To aid in illustrating this process, we have created some dynamic images which are available on request from the authors.

2. Univariate Response

Figure 1 shows a graphical analysis of two simulated calibration data sets from the model

$$Y = \alpha + \beta x + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. In each case a straight line has been fitted to the six "training data" points by least squares, and 95% and 99% prediction intervals have been added (dashed and dotted lines respectively).

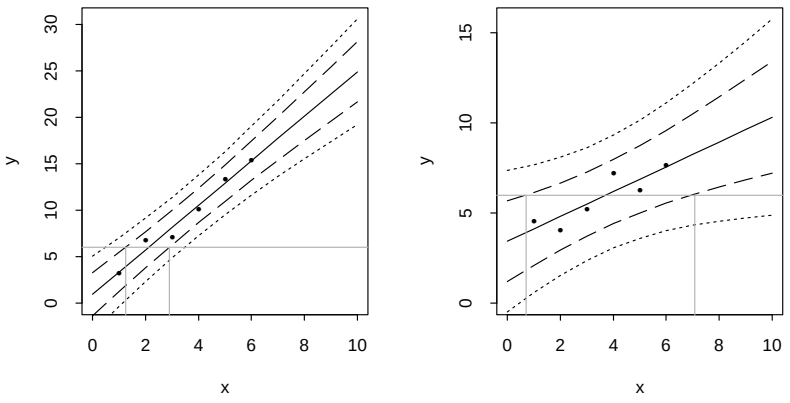

Figure 1: Inverse regression for $Y_0 = 6$ from a simple linear model. The dashed and dotted lines are respectively 95% and 99% prediction intervals. In the second panel the 95% confidence interval for x_0 is finite but the 99$ interval is infinite.

Suppose we now observe $Y_0 = 6$ and want to estimate the corresponding x_0. By adding a horizontal line to the graph, we can read off the x coordinate where this crosses the line of best fit. This is equivalent to using the estimator

$$\hat{x} = \frac{Y_0 - \hat{\alpha}}{\hat{\beta}} \tag{2.1}$$

This we could call the Conditional Least Squares (CLS) estimator since it chooses x_0 to minimize the squared error in Y_0 conditional on the estimated values of parameters α, β, σ. It is also the Maximum Likelihood Estimator (MLE), which we can argue as follows. If we add the new point (x_0, Y_0) to the training data, then for any values of the other parameters we can always place this point on the corresponding line, so that its contribution to the log-likelihood is not a function of α, β. Hence these parameters can be chosen based only on the training data, and $\hat{x}$, $\hat{\alpha}$, $\hat{\beta}$ as defined for CLS will jointly maximize the overall likelihood. This argument still applies if there are multiple unknowns $x_{01}, \ldots, x_{0m}$, and can be extended to the situation of replicates $Y_{01}, \ldots, Y_{0r}$ of the same x_0 (see Brown, 1993 p27), but it will not work for the bivariate response of the next section, or in general if the dimension of Y is greater than the dimension of x. It can also break down in the univariate nonlinear case when there is a horizontal asymptote and the observed Y_0 does not correspond to a fitted value.

It is well-known that a confidence interval for x_0 can be obtained by the process of "inverting a prediction interval" (see for example Carroll and Ruppert, 1988, p56). This is illustrated in both panels of Figure 1, with the 95% confidence limits being given by the points where $y = 6$ crosses the 95% prediction bands. This gives a simple graphical estimation of the uncertainty in x_0, but also illustrates some of the theoretical peculiarities in the use of the CLS estimator. In the second panel we can see that a 99% confidence interval for x_0 would have infinite width, since the line $y = 6$ does not intersect the lower 99% prediction band. The occurrence of this phenomenon is related to the significance of the slope coefficient, which can also be assessed from the graph. If the slope of the regression line is not significantly non-zero then inverse regression seems a dismal prospect. But there will always be some level at which the slope is non-significant. The usually benign assumption of normality leads in this instance to pathological behaviour. This was noted by Hoadley (1970)and resolved using a Bayesian analysis.

Examination of the form of the CLS estimator in equation (2.1) shows that it is related to a Cauchy distribution, being a ratio of normally distributed variables. It does not have a finite mean or variance. This generated considerable controversy in the 1960s, with Krutchkoff (1967) proposing the estimation of x_0 via a regression of x on Y, even for the "controlled calibration" case where the xs in the training data are fixed by design. This can be justified in a Bayesian framework (see Brown, 1993, p98).

This graphical method for obtaining estimates and confidence intervals is particularly useful when the relationship between Y and x is nonlinear or when the errors are heteroscedastic, or both. In cases where a transformation produces approximate linearity and homoscedasticity, it may be advantageous to plot the graph using the untransformed variables for ease of use. Figure 2 shows the PCB data from Bates and Watts (1988), giving ages and PCB concentration in lake trout from Cayuga Lake, NY. They point out that log transformation of Concentration and cube-root of Age seems to produce a straight line with constant variance. In Figure 2 the fitted line and 95% prediction bands have been transformed back to the original scales. The prediction bands give a visual summary of how informative the PCB analysis of a sample would be about the age of the fish from which it is taken. Only at very low concentrations can we get a useful upper limit on the age, but for high concentrations we may be able to say something about the lower age limit. It is unlikely, for example, that a sample with a concentration above 20 ppm would have come from a trout below 8 years old.

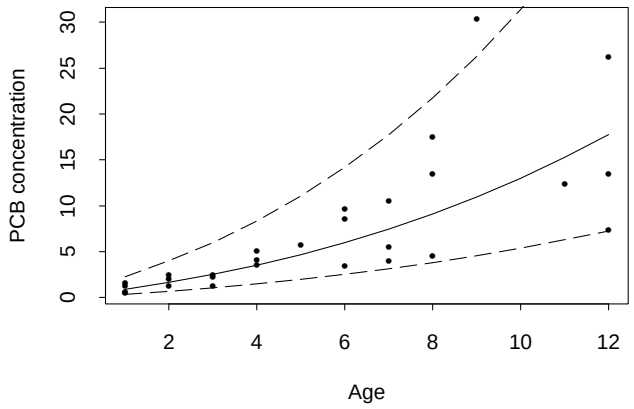

Figure 2: PCB data (Bates and Watts, 1988) with 95% prediction bands.

An important principle of the graphical approach is that once the graph has been produced, it is very easy to use, even though a considerable amount of difficulty may lie in its production. Figure 3 is based on the Toluene data from Rocke and Lorenzato (1995) and uses their two-component model for measurement error

$$Y_i = \alpha + \beta x_i e^{\eta_i} + \epsilon_i \tag{2.2}$$

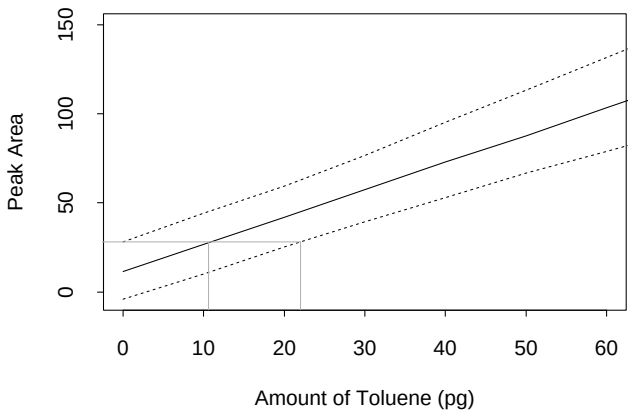

Figure 3: Limits of reliable detection for Toluene data (Rocke and Lorenzato, 1995).

where Y_i is the measured response from a sample with concentration x_i of an analyte. It is assumed that in the absence of error Y_i would be linearly related to x_i. In addition to the usual additive error ϵ_i, this model postulates an additional error term η_i whose effect depends on the concentration level x_i as shown in Equation (2.2). This results in a variance function for Y_i which increases with x_i but does not reduce to zero when $x_i = 0$, thus mimicking the behaviour often observed in analytical measurements. It assumed that both ϵ_i and η_i are normally distributed. One motivation for this model is to assume that random disturbances in the measurement process can be partitioned into those which affect the measurement additively, ϵ_i, and those which have a multiplicative effect, η_i. Further details are given in Rocke and Lorenzato (1995). We focus here on examining the limit of detection of the assay, which measures amounts of toluene by gas chromatography/mass spectrometry (GCMS). The scales of the graph have been restricted to amounts close to zero. Prediction limits could be added using an "exact approximate" method, estimating the parameters by maximum likelihood and using a delta-method approximation. Here we employ instead an "approximate exact" MCMC method suggested by Jones (2004) which accounts properly for the distributions and uncertainty in the estimates, thus giving an accurate picture of the prediction intervals close to zero concentration. To illustrate the limit of detection, we add a horizontal line at the point where the upper prediction band meets $x = 0$; this corresponds to a peak area of 28. Any sample having a GCMS peak area above this value would be regarded as positive for toluene. By adding vertical lines we can see that an amount of 11pg

would have a 50% chance of being detected, whereas 22pg would have a 95% chance. There are a number of competing definitions and calculations for limits of detection and limits of quantitation (see Cox, 2005). Perhaps the best answer would be the graph itself.

3. Bivariate Response

In this section we consider bivariate $\mathbf{Y} = (Y_1, Y_2)'$ and univariate x. An interesting nonlinear example was given by Clarke (1992). Figure 4 plots the lengths of the anterior (Y_1) and posterior (Y_2) horns of a sample of rhinoceros against their known ages in years. The fitted lines are

$$\log Y_i = \alpha + \beta_i \gamma_i^x \qquad i = 1, 2$$

with a common horizontal asymptote α, the parameter values being those given by Clarke (1992). The graph has been plotted on the original, rather than the log scale, so the variation increases considerably with age. Clarke (1992) discusses computational methods for inferring the unknown age of a further animal, adding (x_0, $\mathbf{Y}_0$) to the training data and estimating all parameters simultaneously by maximum likelihood or generalized least squares. Here we examine approximate graphical methods.

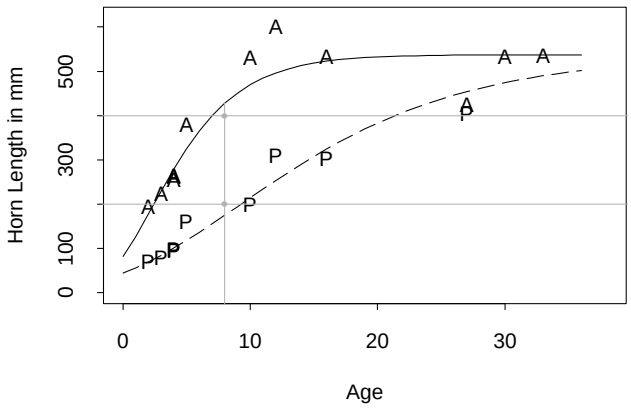

Figure 4: Lengths in mm of the anterior and posterior horns of a sample of 12 rhinocereos, with fitted curves (Clarke, 1992).

Suppose a new rhinoceros presents itself with an anterior horn of 400 mm and posterior horn of 200 mm. Horizontal lines for these two values have been added

to the graph in Figure 4. We now try to add a vertical line which intersects the two horizontal ones at values which seem to fit in with the existing data. This suggests an age of around eight years. A more careful consideration might suggest minimizing in some way the vertical distances of the intersection points from the respective curves. Inspection of the scatter of points around the curves suggests that the anterior horn measurement is more variable. The covariance estimate given by Clarke (1992) confirms this, and also suggests that the departures from the two curves are correlated. Both of these facts argue for moving our estimate towards nine or ten years. From a theoretical perspective we are employing a rough conditional least squares procedure. We could add 95% confidence bands and try to develop a confidence interval for x_0 as in the previous section. Perhaps with a little practice one could imagine moving the vertical line until the intersection points no longer seem to fit in with the existing data.

An alternative viewpoint plots the two horn lengths against each other, as in Figure 5, with age as a parameter tracing out the "response path" as it varies. Such plots have been used before, for example Jones and Rocke (2001) who use multiple response paths for different herbicides to identify and quantitate unknown samples with two immunoassays.

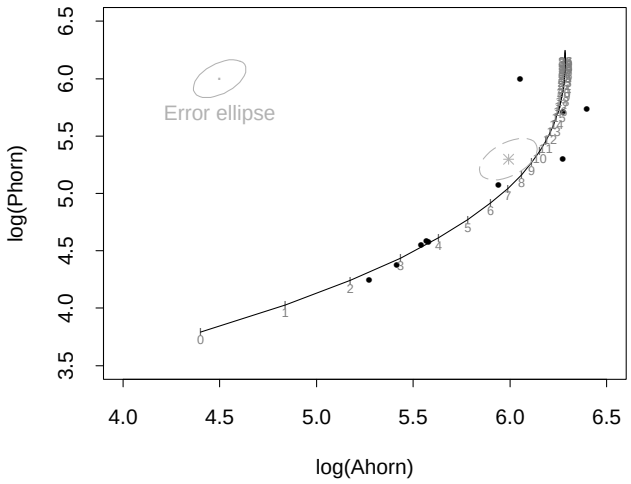

Figure 5: Bivariate response paths for the Rhino data of Figure 4.

A minor innovation here is that the age values have been added as text along the response path. We are using a log scale for both axes since the model assumes constant error covariance on the log scale; however the spread of points around

the response path suggests that the variances are still increasing with age. Our new rhinoceros is plotted as an asterisk at the point $(\log 400, \log 200)$ in Figure 5. Its age is determined by the point on the response path nearest to this new point, that is we project the observed point onto the response path. The direction of this projection should take account of the error covariance, so we have added to the plot a 95% probability ellipse based on the estimated error covariance matrix. We now imagine this ellipse centered on the new point and expanding or contracting until it just touches the response path, at around nine years, as shown in Figure 5. This is a CLS estimator since the response path is regarded as fixed. In contrast to the univariate case, it does not coincide with MLE because the latter would move the estimated response path nearer to the new point. Note that the 95% ellipse does not yield a satisfactory 95% confidence interval by its intersection with the response path; indeed it is possible that there will be no intersection if the new point is far enough away. This situation is discussed by Brown (1993, p89). In the linear case the departure of a new point from the response path can be resolved into orthogonal components along and perpendicular to the path, and a confidence interval developed by considering only the first component. Perhaps a reasonable graphical technique would be, having determined the point of projection, to center the 95% error ellipse on this point and use its intersection with the response path. This suggests an interval estimate of around 8 to 12 years. Maximum likelihood gives an estimated age of 10.1 years, and a profile-likelihood based confidence interval is approximately (7, 14). The extra width of the calculated interval perhaps reflects the uncertainty in the model estimated from quite a small sample.

4. Repeated Measures

If the training data have a hierarchical structure, for example a series of measurements on each of a number of individuals, the model needs to allow different parameters for each individual to reflect population heterogeneity. Such models are variously known as mixed effects, repeated measures, longitudinal or hierarchical models, and have been widely studied (Laird and Ware, 1982; Diggle, Liang and Zeger, 1994; Goldstein, 1995; Davidian and Giltinan, 1995). In the linear case we might have, for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$:

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$$

where $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ and $\beta_i \sim N(\beta, \sigma_\beta^2)$.

Menzefricke (1998) demonstrated the use of MCMC methods for inverse regression in a univariate linear (in parameters) hierarchical model, using as an example the tensile index of paper pulp as a function of beating time. Here we use a bivariate nonlinear example from Jones, Keedwell, Noble and Hedderley

(2005). The motivation for this work was to investigate the possibility of accurately determining the age of a tern chick using easily-obtained measurements of wingspan (Y_1) and weight (Y_2). The model

$$\begin{pmatrix} Y_{1,ij} \\ Y_{2,ij} \end{pmatrix} = \begin{pmatrix} \alpha_i + \beta_i t_{ij} \\ c_i/[1 + dc_i \exp(-b_i t_{ij})] \end{pmatrix} + \begin{pmatrix} \epsilon_{1,ij} \\ \epsilon_{2,ij} \end{pmatrix}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$, was fitted using MCMC with d as a fixed parameter and $(\alpha_i, \beta_i, b_i, c_i)'$ as multivariate normal. We have made some changes to the model of Jones et al (2005), dropping the "slow" group, whose existence was questionable, and assuming a lognormal distribution for the growth rate parameter b_i.

Figure 6 shows the response paths for 150 simulated terns from the fitted model. Age has been shown by changing the color of the response path every two days. Inverse regression for tern chicks of unknown age would be accomplished by plotting the wingspan and weight, and referring the local color(s) to the key. The "braiding" of the response paths produces some mixing of colors and gives an impression of the uncertainty due to the variability in growth parameters. Uncertainty due to the error terms ϵ_1, ϵ_2 can be incorporated by referring to the error ellipse.

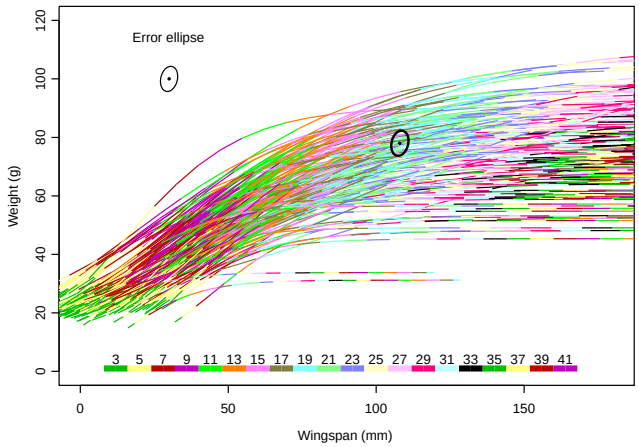

Figure 6: Response paths for growth in wingspan and weight of tern chicks (Jones *et al.*, 2005), with color changing every two days.

Suppose, using an example from Jones et al (2005), that a new chick is observed with wing length 108 mm and weight 78 g. The color at this point on the graph gives the approximate age of the chick. If an error ellipse is drawn centered

at this point, it gives an impression of the uncertainty in the estimate. In our case, the color is mostly that of 19 and 21 days. The Bayesian analysis given by Jones et al (2005) gives a 95% credible interval of (19.7, 21.6).

In constructing such a graph, there are a number of parameters to consider which can have an important bearing on the useability of the result: the number of simulated paths, the thickness of the lines, the rate at which the color changes, and the colors themselves. If the colors change too rapidly (for example, every day in the tern chick graph) there is too much mixing of colors; if they change too slowly, precision is lost. To aid in exploring the effects of these parameters on visualization of the information and in optimizing the design of the graph, we have created some software which can be obtained from the authors on request. The software allows the user to specify the band width – the range of ages represented by a single color on the display. It then calculates the number of colors required (age range/bandwidth), and allocates them to age ranges in hue sequence, with saturation and lightness both alternating between high and low. That is, increasing ages are represented by colors in color-wheel order, but dull, dark colors alternate with light, bright colors. This alternation between high and low values of saturation and lightness ensure that adjacent colors on the graph are clearly distinguishable.

5. Discussion

We have summarized and illustrated the ways in which graphical methods can aid in visualizing the process of inverse regression, and can be used to derive estimates when only an approximate value is required. Some computational issues have not been carefully addressed. We have focussed mostly on conditional least squares – although the uncertainty in the model estimates is incorporated in the inversion of prediction intervals in our univariate examples, it is not allowed for in the bivariate cases. Generally, as the situation becomes more complex the approximation becomes cruder. In the rhinoceros example it would be much easier and more intuitive to take the point on the response path nearest to the new data. Allowing for the covariance of the errors by using a Mahalonobis distance is more accurate but much more difficult, although the use of an error ellipse makes it possible. In examining the colors inside the error ellipse of the tern chick graph, more weight should be given to points nearer the center of the ellipse, but also to points on response paths closer to the average path. Again this detracts from the intuitive appeal and ease of use of the graphical method.

We have not distinguished here between single use and multiple use of calibration curves. Computational methodologies can be adapted for making joint probability statements about multiple inverse estimates. We have also not considered the case of uncontrolled calibration, where the distribution of x values in

the training data is informative about the unknown x_0. Both of these issues are discussed by Brown (1993).

We have introduced a new graphic for illustrating bivariate growth patterns and performing approximate age estimation. This could easily be extended to situations in which different groups have different growth characteristics, so that discrimination as well as inverse regression would be required. Good design is always important for producing useful graphs, but it is of particular importance in this case. The choice of colors requires a combination of theory (the color-wheel) and trial-and-error if an adequate visualization is to be achieved. Good design can be facilitated by software that allows the parameters to be changed interactively before a final version is printed.

Acknowledgments

We wish to thank an anonymous referee for useful suggestions and references.

References

Brown, P. J. (1993). *Measurement, Regression and Calibration.* Oxford University Press.

Bates, D. M., and Watts, D. G. (1988). *Nonlinear Regression and its Applications.* Wiley.

Carroll, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall.

Clarke, G. P. Y. (1992). Inverse estimates from a multiresponse model. *Biometrics* **48**, 1081-1094.

Cox, C. (2005). Limits of quantitation for laboratory assays. *Applied Statistics* **54**, 63-78.

Davidian, M., and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall.

Diggle,A., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Clarendon Press.

Hoadley, B. (1970). A Bayesian look at inverse linear regression. *Journal of the American Statistical Association* **65**, 356-369.

Goldstein, H. (1995). *Multilevel Statistical Models.* Halsted Press.

Laird, N. M., and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Jones, G. (2004). Markov chain Monte Carlo estimation for the two-component model. *Technometrics* **46**, 99-107.

Jones, G., Keedwell, R. J., Noble, A. D. L. and Hedderley, D. I. (2005). Dating chicks: Calibration and discrimination in a nonlinear multivariate hierarchical growth model. *Journal of Agricultural, Biological and Environmental Statistics* **10**, 306-320.

Jones, G. and Rocke, D. M. (2001). Analyte identification in multivariate calibration. *Biometrics* **57**, 571-576.

Menzefricke, U. (1998). Bayesian prediction in growth-curve models with correlated errors. *Test* **8**, 75-93.

Krutchkoff, R. G. (1967). Classical and inverse methods of calibration. *Technometrics* **9**, 5425-439.

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

Received January 30, 2007; accepted August 20, 2007.

Geoffrey Jones
Computer Science
Institute of Information Sciences and Technology
Massey University
Palmerston North, New Zealand
g.jones@massey.ac.nz

Paul Lyons
Computer Science
Institute of Information Sciences and Technology
Massey University
Palmerston North, New Zealand
p.lyons@massey.ac.nz