Image Cover Sheet

CLASSIFICATION	SYSTEM NUMBER 509992
UNCLASSIFIED	

TITLE

A LEAST-SQUARES STRAIGHT-LINE FITTING ALGORITHM WITH AUTOMATIC ERROR DETERMINATION

System Number:

Patron Number:

Requester:

Notes:

DSIS Use only:

Deliver to:

National Defence
Défense nationale

A LEAST-SQUARES STRAIGHT-LINE FITTING ALGORITHM WITH AUTOMATIC ERROR DETERMINATION

by

T. Thayaparan and W.K. Hocking

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 98-004

Canada

June 1998
Ottawa

National Defence | Défense nationale

A LEAST-SQUARES STRAIGHT-LINE FITTING ALGORITHM WITH AUTOMATIC ERROR DETERMINATION

by

T. Thayaparan
Ground Based Radar Group
Surface Radar Section

and

W.K. Hocking
University of Western Ontario

DEFENCE RESEARCH ESTABLISHMENT OTTAWA
TECHNICAL NOTE 98-004

Project
05EE17

June 1998
Ottawa

Abstract

We demonstrate a new generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets which are expected to be linearly correlated are subject to different uncertainties in their measurements. The algorithm determines not only the optimum slope, but also produces estimates of the intrinsic errors associated with each data set. It requires almost no initial information about the errors in each data set.

Executive Summary

One of the problems that frequently confronts an experimentalist is that of fitting a straight line to data, that is, the problem of determining the functional relationship between two variables. A least-squares calculation is the most common approach to this problem and, in its simplest form, one variable is subject to error while the other is error-free. This procedure is described in almost all elementary books on statistics.

A slightly more complex case is that both variables have errors, which are well known. Considerable work has been done on fitting procedures which take into account errors in both variables. These fitting procedures generally work very well for the tasks for which they are assigned. However, there are times when two data sets are compared in which the intrinsic measurement errors of one or both variables are unknown. In this case, there is no algorithm available to determine the best-fit line. The main objective of this paper is to demonstrate a new procedure for this objective.

In this paper, we demonstrate a new generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets which are expected to be linearly correlated are subject to different uncertainties in their measurements. The algorithm determines not only the optimum slope, but also produces estimates of the intrinsic errors associated with each data set. It requires almost no initial information about the errors in each data set. The algorithm works best if the user knows which of the two variables has the smallest intrinsic error, but is not limited to this condition. We hope that this process will find application in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements.

Contents

List of Figures

List of Tables

1 Introduction

Least-squares techniques for examining best-fit lines to correlated data sets are well established, but in general are restricted to cases in which the errors in the two data sets being compared are known. The simplest case is that in which there is no error in one variable (usually plotted as the abscissa), and the error in the other variable is known. This procedure is described in almost all elementary books on statistics [e.g., *Young,* 1962; *Bevington,* 1969; *Daniel and Wood,* 1971; *Brandt,* 1976; *Taylor,* 1982].

A slightly more complex case is that in which both variables have errors, but both errors are well known. Examples of such process have been shown by (amongst others) *York* [1966], *Barker* [1974], *Orear* [1982], *Lybanon* [1984], *Miller and Dunn* [1988], *Reed* [1989, 1992], and *Jolivette* [1993]. These fitting procedures generally work very well for the tasks for which they are assigned.

However, there are times when two data sets are compared in which the intrinsic measurement errors of one or both variables are unknown. In this case, there is no algorithm available to determine the best-fit line. The purpose of this paper is to demonstrate a new procedure for this objective.

This paper is broken up as follows. We begin by simulating some data using computer techniques, in which the intrinsic errors are pre-specified. We then demonstrate the effects of employing existing methods of data analysis to these data sets, and point out the limitations of these procedures. We then demonstrate how these standard processes lead to our new procedure, and develop our new algorithm. Finally, we use new artificially generated data to demonstrate the application of our new method.

2 Standard analysis procedures

As an initial study, we used computer-generated data to re-visit standard least-squares fitting procedures. Our model works as follows. We generate two data sets, one representing the abscissa (x_i), and the other the ordinate (y_i). These are generated by randomly selecting values x_i from a Gaussian distribution with a pre-specified standard deviation. In our first case, we will demonstrate the situation for a standard deviation of $\Sigma_x = 40$ and $\Sigma_y = 40$. We then calculated y_i values using the expression

$$y_i = g_o\, x_i + c_1 \tag{1}$$

For our initial calculations, we have chosen $g_o = 1$, which produces a standard deviation in the y_i values (Σ_y) equal to Σ_x. Notice that these Σ values are *not* errors, but represent the spread in our raw data points. Other options are considered later. This process then gave a straight line graph. Our next step was to partly randomize the points. For each point, we added a random number to the x_i value, and a different random number to the y_i value, where each randomly generated number was derived from a Gaussian distribution with standard deviations σ_x and σ_y respectively. We generated about 300 points per such realization in the first instance.

Our next step was to perform various types of least-squares fitting procedures. We removed the mean $\bar{x}$ value from the x_i values, and the mean $\bar{y}$ value from the y_i values, and then applied standard fitting procedures in two different ways. Firstly, we fitted the data assuming that all the x_i values have no error. This gave an equation which we denote as

$$y_i = g_x\, x_i + c_2 \tag{2}$$

Because the data were normally distributed, as were the errors, we found that c_2 was generally close to zero. We will therefore not further investigate this value, and will concentrate on the parameter g_x. For real data, we assume that the user will have already removed the means before applying our method.

Next we fitted the same data assuming that the y_i values have no error, and obtained lines of the type

$$x_i = (1/g_y)\, y_i + c_3 \tag{3}$$

or

$$y_i = g_y\, x_i + c_4 \tag{4}$$

As a final check, we also fitted our data using the algorithm described by York [1966], in which we used our known values for σ_x and σ_y to estimate the mean slope. We will denote this

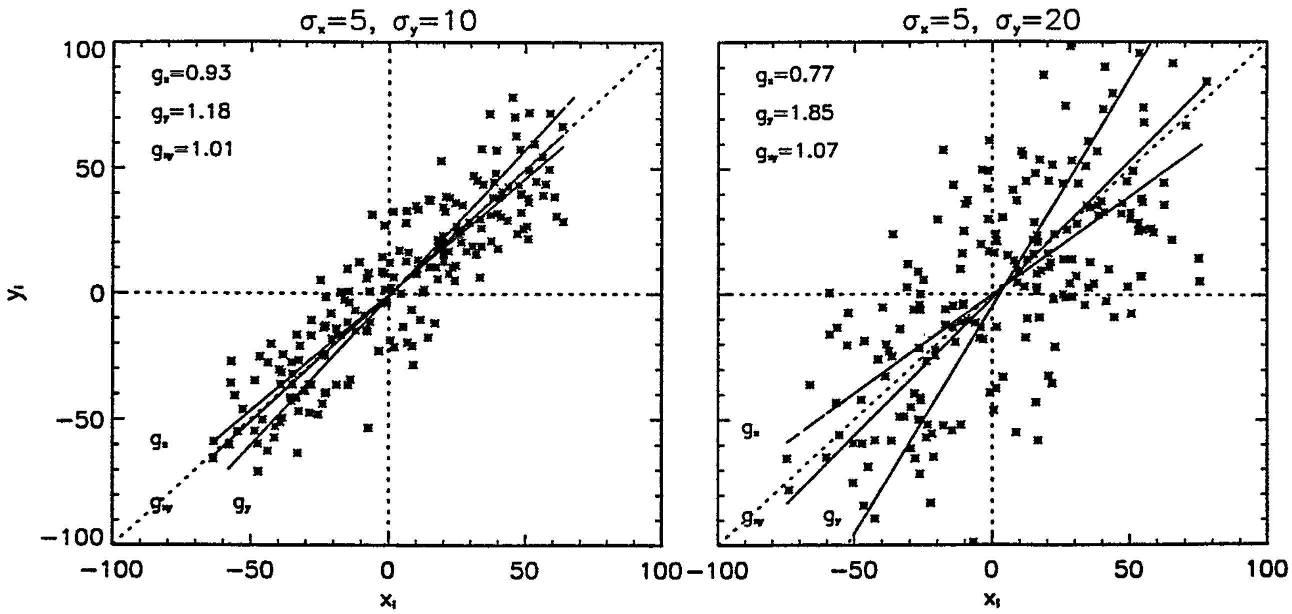

Figure 1: Examples of the slopes obtained from three different methods applied to some sample data. The symbols g_x and g_y refer to the slopes obtained from the regressions y on x and x on y respectively, while g_{xy} refers to the slope obtained from the algorithm described by *York* [1966]. σ_x and σ_y are measurement errors in x and y respectively.

estimate as g_{xy}. This latter step is not required, but was simply done as a consistency check. We then repeated this process for each (σ_x, σ_y) pair for many different realizations (typically 300). Figure 1 shows examples of our fitting attempts for cases with $(\sigma_x = 5, \sigma_y = 10)$ and $(\sigma_x = 5, \sigma_y = 20)$.

As noted, we have repeated such fits for many hundreds of different realizations, and for different combinations of σ_x and σ_y, and different numbers of points per sample. The results of the mean slopes g_x, g_y, and g_{xy}, and associated standard deviations for the mean (the latter being denoted as Δg_x, Δg_y, and Δg_{xy}) are shown as contour plots in Figure 2. It is clear that the York [1966] algorithm has worked well (bottom panel), with all slope estimates being close to unity (the original slope). We will therefore not further discuss the results of the York algorithm.

However, the distributions of g_x and g_y are striking in the way that the contours are aligned almost parallel to the axes (see Figure 2). It is this rather special alignment which permits us to develop the algorithm described in this paper. The arrangement of these lines shows

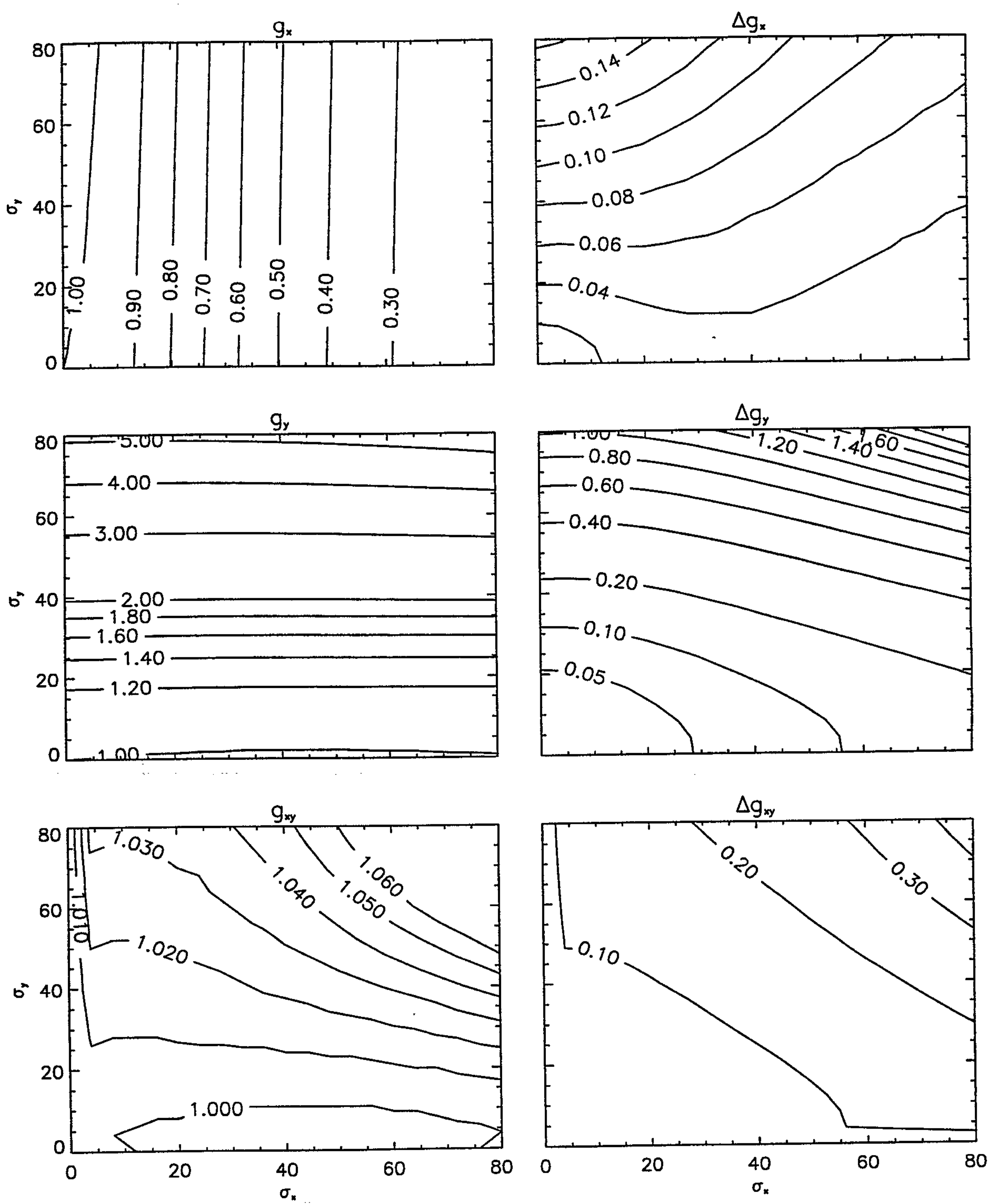

Figure 2: Slopes obtained from three different methods as a function σ_x and σ_y. g_x and g_y refer to the slopes obtained from the regressions y on x and x on y respectively. g_{xy} refers to the slope obtained from the algorithm described by *York* [1966]. σ_x and σ_y are intrinsic noise and measurement errors in x and y respectively. Δg_x, Δg_y, and Δg_{xy} are the standard deviations for the means (standard error) of these various slopes.

that the estimate of g_x, although erroneous, is pretty much independent of σ_y. Likewise, g_y depends very weakly on σ_x. Thus we may write

$$g_x \approx g_x(\sigma_x) \quad ; \quad g_y \approx g_y(\sigma_y) \tag{5}$$

There are slight deviations from this law for small σ_x and large σ_y in the top graph, but otherwise these dependencies are fairly true. Because of this striking character, we then decided to develop an analytical expression for g_x as a function of σ_x. Our results are shown in Figure 3, where we plot g_x as a function of σ_x for selected σ_x. We have actually taken the case $\sigma_y = 0$, but as noted g_x depends only very weakly on σ_y. We have also re-plotted our ordinate and abscissa as σ_x/Σ_x and g_x/g_o. We have done this because we expect that these scalings should apply, and we have confirmed this by repeating our analyses using different values of g_o and Σ_x. Henceforth we will generally work with such normalized values.

At this point, we need to indicate an important assumption concerning the sign of the slope. For our preceding examples, we used cases for which the best fit lines between data sets had a positive slope. This may seem restrictive, but in fact it is not, as will now be seen. In any realistic situation in which the data show reasonable correlation, g_x and g_y will have the same sign, so it is easily determined what the sign of the true best fit-slope, g_o, should be. If it is negative, we can simply reverse the signs of all of the y_i values, and the best fit line will then have positive slope. We may then apply the algorithm, and upon its completion (and determination of the best fit slope $\tilde{g_o}$), we may then get our true best-fit slope by reversing the sign of $\tilde{g_o}$. Thus our restriction to positively correlated data is not a limitation of the technique, and the rest of the discussion will proceed assuming such positively correlated data.

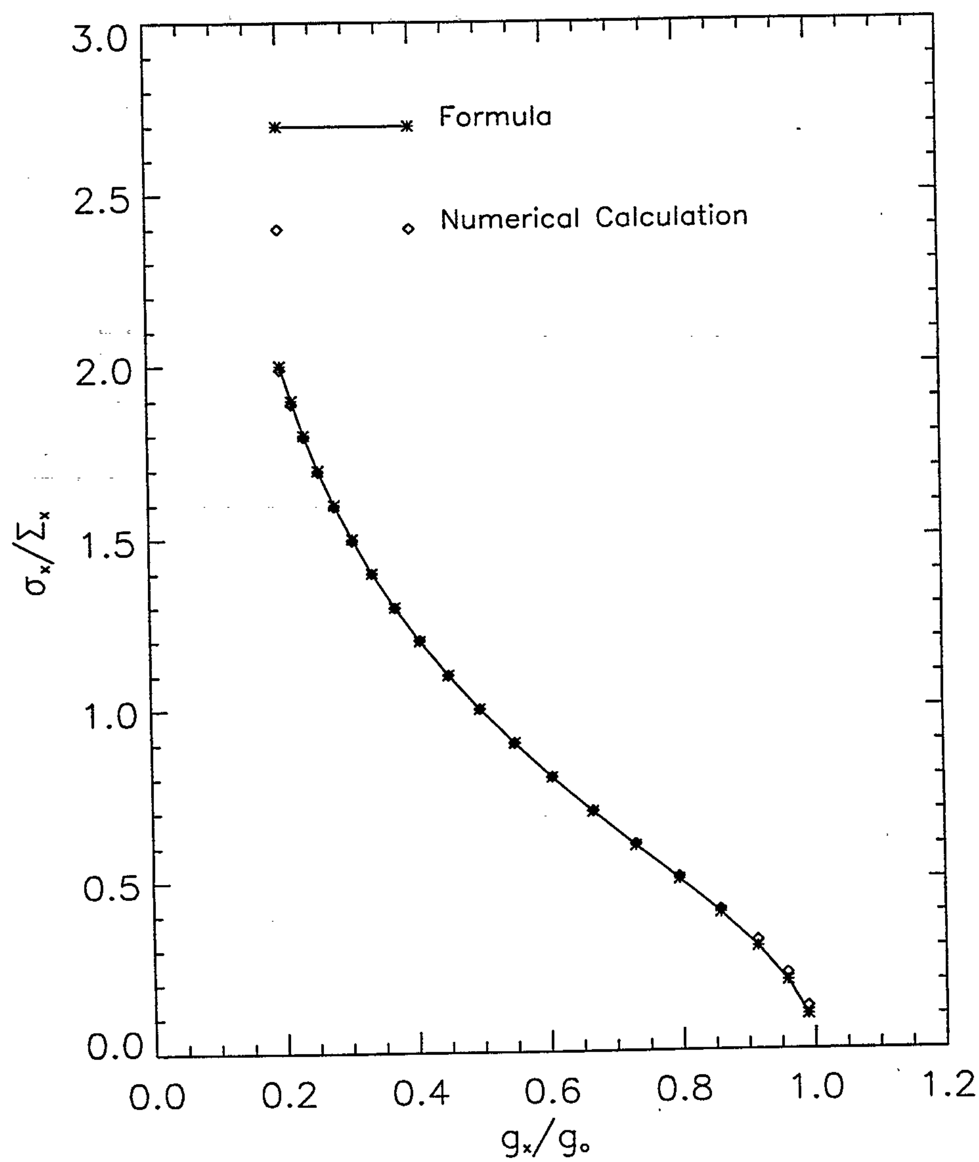

Figure 3: Variation of σ_x/Σ_x as a function of g_x/g_o. The solid line with asterisk symbols is the empirical formula given by equation (6), and the diamond symbols are the values calculated from direct application of the numerical model described in Section 2.

We may now return to our developmental discussion. Following generation of Figure 3, we determined an analytical function which fitted these points. The curve is continuous and continuously differentiable, and is plotted as the solid line in the figure. The expression we have used is complicated, but quite accurate, and is given by

$$\begin{aligned}\frac{\sigma_x}{2.15\,\Sigma_x} = (1-\frac{g_x}{g_o}) + (1-\frac{g_x}{g_o})^9 - 0.049\,\sin[3.391\,(\frac{g_x}{g_o}+0.255)^2] \\ +0.01\,\exp[\frac{-(\frac{g_x}{g_o}-1)^2}{2(0.1)^2}] + 0.007\,\exp[\frac{-(\frac{g_x}{g_o}-0.58)^2}{2(0.15)^2}] \\ +0.012\,\exp[\frac{-(\frac{g_x}{g_o}-0.94)^2}{2(0.022)^2}] + 0.025\,\exp[\frac{-(\frac{g_x}{g_o}-2)}{0.1}] \\ +0.01\,\exp[\frac{(\frac{g_x}{g_o}-1)}{0.05}]\,\cos[\frac{(1-\frac{g_x}{g_o})}{0.1}+\frac{\pi}{6}] \qquad (6)\end{aligned}$$

A more succinct formula may be found in due course, but for now our purpose is just to demonstrate our method. Furthermore, we also recognize that the relation between g_o/g_y and σ_y obeys the same expression (although note that we use g_o/g_y here compared to g_x/g_o above).

$$\begin{aligned}\frac{2.15\,\sigma_y}{\Sigma_y} = (1-\frac{g_o}{g_y}) + (1-\frac{g_o}{g_y})^9 - 0.049\,\sin[3.391\,(\frac{g_o}{g_y}+0.255)^2] \\ +0.01\,\exp[\frac{-(\frac{g_o}{g_y}-1)^2}{2(0.1)^2}] + 0.007\,\exp[\frac{-(\frac{g_o}{g_y}-0.58)^2}{2(0.15)^2}] \\ +0.012\,\exp[\frac{-(\frac{g_o}{g_y}-0.94)^2}{2(0.022)^2}] + 0.025\,\exp[\frac{-(\frac{g_o}{g_y}-2)}{0.1}] \\ +0.01\,\exp[\frac{(\frac{g_o}{g_y}-1)}{0.05}]\,\cos[\frac{(1-\frac{g_o}{g_y})}{0.1}+\frac{\pi}{6}] \qquad (7)\end{aligned}$$

i.e., as above with g_x/g_o replaced by $(g_y/g_o)^{-1}$.

This now means that if we have a set of points (x_i, y_i), and we know the intrinsic error in one of these variables, we can determine uniquely the "true" best fit line $\tilde{g}_o$ and also the intrinsic error in the other variable. For example, suppose the user knows σ_x. Then, one can calculate the experimental standard deviations ς_x and ς_y for the two data sets. These standard deviations include both the natural spread of the data and the intrinsic errors. Then we may determine an estimator for Σ_x, which we will denote as $\tilde{\Sigma}_x$, through the relation

$$\tilde{\Sigma}_x = \sqrt{\varsigma_x^2 - \sigma_x^2} \tag{8}$$

Next, the user applies standard fitting procedures, first assuming $\sigma_x = 0$ to get g_x, and then $\sigma_y = 0$ to get g_y. One may use Figure 3 or equation (6) to determine an estimator for g_x/g_o (which we will denote as $\tilde{g_{xo}}$) from the knowledge of $\sigma_x/\tilde{\Sigma}_x$. Thus, knowing g_x, and g_x/g_o, one can readily determine $\tilde{g}_o$, our estimate of g_o. Furthermore, now that $\tilde{g}_o$ is known, the user can determine an estimator for g_y/g_o (which we will denote as $\tilde{g_{yo}}$). Further application of equation (7) then permits determination of an estimator for σ_y/Σ_y. This may in turn be used to determine $\tilde{\Sigma}_y$ through the relation

$$\tilde{\Sigma}_y = \varsigma_y / \sqrt{1 + (\frac{\tilde{\sigma}_y}{\tilde{\Sigma}_y})^2} \tag{9}$$

Since $\varsigma_y = \sqrt{\Sigma_y^2 + \sigma_y^2}$, an estimator for σ_y can be uniquely determined as well. Hence given σ_x, the user can determine the slope of the best-fit line, $\tilde{g}_o$, and also determine $\tilde{\sigma}_y$. This is useful, but it is not the full story.

The above description assumes that the user knows the intrinsic error σ_x. But what if this is not known ? Can we determine it ?

Suppose we "guess" σ_x, and carry out the above process. Then we can obtain $\tilde{g}_o$, $\tilde{\Sigma}_x$, $\tilde{\Sigma}_y$, and $\tilde{\sigma}_y$ for this proposed value of σ_x. However, we do not know if we have chosen the correct value for σ_x. This dilemma can be solved, because there is one extra piece of information available which has not yet been used. The means have been removed, so that if there were no random errors, we would expect Σ_y/Σ_x should equal g_o. Hence we may say that, to quite high accuracy,

$$\tilde{\Sigma}_y/\tilde{\Sigma}_x \simeq \tilde{g}_o \tag{10}$$

It is essential that this condition is closely satisfied in order for our selection of σ_x to be considered as acceptable. We may thus use this as our final diagnostic test. If we try different values of σ_x, and repeat the above process, we can examine the ratio $(\tilde{\Sigma}_y/\tilde{\Sigma}_x)/\tilde{g}_o$, and choose the value of σ_x which allows this parameter to be closest to 1. Alternatively, we may find the value where the parameter $\Xi = [((\tilde{\Sigma}_y/\tilde{\Sigma}_x)/\tilde{g}_o) - 1]$ is closest to zero.

A.		Remove the means from both data sets.
B.		Do standard least-squares fit to obtain g_x and g_y.
C.		If g_x and g_y are negative, reverse the signs on all y_i values, and the signs of g_x and g_y.
D.		Find experimental standard deviations of the data sets, ς_x and ς_y. (Note these include both natural spread and intrinsic errors).
E.		Determine appropriate steps for σ_x, (let these step sizes be denoted as $\Delta\sigma_x$).
F.		Step σ_x from zero, upward in steps of $\Delta\sigma_x$.
G.		For each value of σ_x, do the following.
	G1.	Find $\tilde{\Sigma}_x = \sqrt{\varsigma_x^2 - \sigma_x^2}$.
	G2.	Find $\sigma_x/\tilde{\Sigma}_x$, and then use equation (6) to determine an estimate for g_x/g_o (call it $\tilde{g_{xo}}$).
	G3.	From $\tilde{g_{xo}}$, and from known value of g_x, determine $\tilde{g}_o$ (an estimate of g_o).
	G4.	From $\tilde{g}_o$, and the previously determined value of g_y, determine $\tilde{g_{yo}} = g_y/\tilde{g}_o$.
	G5.	Determine $\tilde{R}_y = \tilde{\sigma}_y/\tilde{\Sigma}_y$ from equation (7).
	G6.	Determine $\tilde{\Sigma}_y$ from $\tilde{\Sigma}_y = \varsigma_y/\sqrt{1 + \tilde{R}_y^2}$.
	G7.	Determine $\tilde{\sigma}_y = \sqrt{\varsigma_y^2 - \tilde{\Sigma}_y^2}$.
	G8.	Now calculate $\Xi = ((\tilde{\Sigma}_y/\tilde{\Sigma}_x)/\tilde{g}_o) - 1$.
	G9.	Go to G1, and repeat the loop G1 to G8 and search for the point where Ξ is closest to zero.
H.		If a sign reversal was required in step C, reverse the signs of all the y_i values again, and reverse the sign of the resulting slope $\tilde{g}_o$.
I.		Add the means back to the resulting equation.
		Note: We recommend choosing the $\{x_i\}$ and $\{y_i\}$ values so that $\sigma_x \leq \sigma_y$.

Table 1: Procedure for unique determination of g_o, σ_x, and σ_y.

Thus repeated application of successive values of σ_x, coupled with the above test for $\Xi = 0$, permit us to determine the best estimator for σ_x. Hence this iterative approach allows us to uniquely determine estimates for g_o, σ_x, and σ_y. No prior knowledge about any of these parameters is required, although we will show shortly that from a pragmatic perspective it is wisest to choose the variables $\{x_i\}$ and $\{y_i\}$ such that $\sigma_x \leq \sigma_y$. Our basic procedure is summarized in Table 1.

$\tilde{\sigma_x}$	$\tilde{\Sigma_x}$	ς_x	$\tilde{\sigma_y}$	$\tilde{\Sigma_y}$	ς_y	$\tilde{g_x}$	$\tilde{g_y}$	$\tilde{g_o}$	$\tilde{\Sigma_y}/\tilde{\Sigma_x}$	Ξ
0.50	38.62	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.984	16.9
1.00	38.61	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.984	16.6
1.50	38.59	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.984	16.2
2.00	38.57	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.985	15.6
2.50	38.54	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.985	14.8
3.00	38.50	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.986	13.9
3.50	38.46	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.987	12.8
4.00	38.41	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.989	11.5
4.50	38.36	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.990	10.0
5.00	38.30	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.992	8.4
5.50	38.23	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.994	6.6
6.00	38.15	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.996	4.6
6.50	38.07	38.62	16.43	37.98	41.38	0.96	1.19	1.000	0.998	2.4
7.00	37.98	38.62	16.39	38.00	41.38	0.96	1.19	1.001	1.000	1.4
7.50	37.89	38.62	16.20	38.08	41.38	0.96	1.19	1.005	1.005	0.3
8.00	*37.78*	*38.62*	*15.99*	*38.17*	*41.38*	*0.96*	*1.19*	*1.009*	*1.010*	*0.0*
8.50	37.67	38.62	15.76	38.26	41.38	0.96	1.19	1.014	1.016	-0.8
9.00	37.56	38.62	15.51	38.36	41.38	0.96	1.19	1.019	1.021	-1.9
9.50	37.43	38.62	15.22	38.48	41.38	0.96	1.19	1.025	1.028	-2.8
10.00	37.30	38.62	14.87	38.62	41.38	0.96	1.19	1.032	1.035	-3.1
10.50	37.17	38.62	14.43	38.78	41.38	0.96	1.19	1.040	1.044	-3.0
11.00	37.02	38.62	13.87	38.99	41.38	0.96	1.19	1.051	1.053	-2.0
11.50	36.87	38.62	13.26	39.20	41.38	0.96	1.19	1.063	1.063	-2.6
12.00	36.71	38.62	12.70	39.38	41.38	0.96	1.19	1.074	1.073	-2.8
12.50	36.54	38.62	12.20	39.54	41.38	0.96	1.19	1.084	1.082	-3.1
13.00	36.37	38.62	11.73	39.68	41.38	0.96	1.19	1.095	1.091	-3.3
13.50	36.18	38.62	11.21	39.83	41.38	0.96	1.19	1.105	1.101	-3.6
14.00	35.99	38.62	10.53	40.02	41.38	0.96	1.19	1.117	1.112	-4.2
14.50	35.80	38.62	9.59	40.25	41.38	0.96	1.19	1.128	1.125	-4.3
15.00	35.59	38.62	8.36	40.53	41.38	0.96	1.19	1.141	1.139	-4.6

Table 2: Sequence of calculations for the parameter Ξ when input values are $\Sigma_x = 40$, $\sigma_x = 8$, $\Sigma_y = 40$, and $\sigma_y = 16$.

3 Tests of our models

Having established our numerical iterative model for $\tilde{g_o}$, $\tilde{\sigma_x}$, and $\tilde{\sigma_y}$, it is of course necessary to test its accuracy. For this, we again turn to simulation.

We repeated the process described in section 2 to generate our original $\{x_i\}$ and $\{y_i\}$ values, but this time for a wide variety of values for g_o, σ_x, and σ_y. Then, for each data set, we applied the algorithm described in Table 1. Table 2 shows a sequence of calculations for σ_x in steps of 0.5 (un-normalized units). The important parameter to observe is Ξ, as written in the last column of the table. In this particular example, we see that Ξ passes through zero

$\tilde{\sigma}_x$	$\tilde{\Sigma}_x$	ς_x	$\tilde{\sigma}_y$	$\tilde{\Sigma}_y$	ς_y	$\tilde{g}_x$	$\tilde{g}_y$	$\tilde{g}_o$	$\tilde{\Sigma}_y/\tilde{\Sigma}_x$	Ξ
0.50	40.99	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.944	28.8
1.00	40.98	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.945	28.6
1.50	40.97	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.945	28.2
2.00	40.95	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.946	27.6
2.50	40.92	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.948	27.0
3.00	40.88	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.949	26.1
3.50	40.84	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.951	25.1
4.00	40.80	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.953	24.0
4.50	40.75	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.956	22.7
5.00	40.69	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.959	21.2
5.50	40.62	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.962	19.6
6.00	40.55	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.965	17.8
6.50	40.48	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.969	15.9
7.00	40.39	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.973	13.8
7.50	40.30	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.977	11.5
8.00	40.21	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.982	9.1
8.50	40.10	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.987	6.5
9.00	39.99	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.993	3.8
9.50	39.88	40.99	21.39	79.69	82.51	1.89	2.14	2.000	1.998	0.9
10.00	*39.76*	*40.99*	*20.72*	*79.87*	*82.51*	*1.89*	*2.14*	*2.009*	*2.009*	*0.1*
10.50	39.63	40.99	19.63	80.14	82.51	1.89	2.14	2.021	2.022	-0.7
11.00	39.49	40.99	18.11	80.50	82.51	1.89	2.14	2.036	2.038	-1.2
11.50	39.35	40.99	15.97	80.95	82.51	1.89	2.14	2.054	2.057	-1.4
12.00	39.20	40.99	13.36	81.42	82.51	1.89	2.14	2.076	2.077	-1.8
12.50	39.04	40.99	10.75	81.81	82.51	1.89	2.14	2.097	2.095	-2.6
13.00	38.88	40.99	8.31	82.09	82.51	1.89	2.14	2.117	2.112	-2.8
13.50	38.71	40.99	5.93	82.30	82.51	1.89	2.14	2.136	2.126	-4.6
14.00	38.53	40.99	3.50	82.44	82.51	1.89	2.14	2.155	2.140	-7.4
14.50	38.34	40.99	0.89	82.51	82.51	1.89	2.14	2.175	2.152	-10.9
15.00	38.15	40.99	1.96	82.49	82.51	1.89	2.14	2.196	2.162	-15.7

Table 3: Sequence of calculations for the parameter Ξ when input values are $\Sigma_x = 40$, $\sigma_x = 10$, $\Sigma_y = 80$, and $\sigma_y = 20$.

at $\sigma_x = 8.0$ (highlighted in italics). We see that our program estimates σ_y to be 15.99 and g_o to be 1.009. Our estimates are very close to our original input values $\sigma_x = 8.0$, $\sigma_y = 16.0$, and $g_o = 1.0$. The fact that our best estimate for g_o is about 1% higher than the true value is a concern, but not a major one. There are clearly cases which give a best fit $\tilde{g_o}$ which is better than our final estimate, especially at low $\tilde{\sigma_x}$ and high $\tilde{\sigma_y}$, but these have erroneous estimates of σ_x and σ_y. Our final estimate gives the best possible simultaneous combination of estimates for g_o, σ_x, and σ_y, and this is the most important point. These small errors probably arise in part because the equation (5) are not exactly true.

Table 3 shows another example of a different realization and alternative input values. In this case, Ξ passes through zero at $\sigma_x = 10.0$ (highlighted in italics) when the corresponding σ_y and g_o values are 20.72 and 2.009 respectively. These estimates are again very close to our original input values $\sigma_x = 10.0$, $\sigma_y = 20.0$, and $g_o = 2.0$. Tables 2 and 3 clearly demonstrate that our algorithm determines not only the optimum slope, but also produces estimates of the intrinsic errors associated with each data set. Of course in a realistic application of our algorithm, we do not need to produce tables like those just shown for every situation. It is adequate to let the computer search for the zero crossing point of Ξ, and this is what we normally do.

We have repeated this procedure for many different combinations of σ_x, σ_y, and g_o, and for many different realizations. Figure 4 shows a collective graph of the $\tilde{\sigma_x}/\tilde{\Sigma_x}$ and $\tilde{\sigma_y}/\tilde{\Sigma_y}$ values which we have estimated compared to the original input values for many different realizations. The relationship is clearly quite linear and confirms that our algorithm is making sensible predictions.

Figure 5 shows differences between our estimate of $\tilde{g_o}$ relative to g_o after subtraction of unity. We clearly produce accuracy of $\sim 0.1\%$ to 1% for the slope of the line. Thus we have demonstrated the validity of our technique, at least for the case in which the raw data and the errors are normally distributed.

We have also repeated our analyses for different numbers of points within any one data set. We have found that provided there are greater than 50 points per data set, we produce error estimators for $\tilde{\sigma_x}$ and $\tilde{\sigma_y}$ which have absolute errors ($\delta\tilde{\sigma_x}$ and $\delta\tilde{\sigma_y}$) of less than about 2.5% of Σ_x and Σ_y respectively (see Figure 4; the spread in $\tilde{\sigma_x}/\tilde{\Sigma_x}$ and $\tilde{\sigma_y}/\tilde{\Sigma_y}$ is about 0.05). However, when lesser numbers of points are used per realization, the performance degrades considerably. Data sets of 20 points become substantially less reliable with regard to estimating σ_x and σ_y. We recommend that this procedure should be restricted to cases in which there are at least 30 points, and preferably 50 or more.

We should finish with a final comment about multiple minima. As noted, we have created tables like Table 2 for many combinations of g_o, σ_x, and σ_y (although, as we have said, in any practical situation we allow the computer program to find the zero points of Ξ). We have noted that for some cases there are multiple zero-crossings, especially for cases of large σ_y. This problem is worse if $\sigma_x > \sigma_y$, so we recommend that if users have any a-priori knowledge about their data, they should choose the $\{x_i\}$ values as those with the smaller σ_x. Determinations should be stopped when $\tilde{\sigma_x}$ exceeds $\tilde{\sigma_y}$, since this very often removes any other zero-crossing cases. If no a-priori knowledge is available, both combinations should be attempted (i.e., first use data set 1 as $\{x_i\}$, and then use data set 2 as $\{x_i\}$).

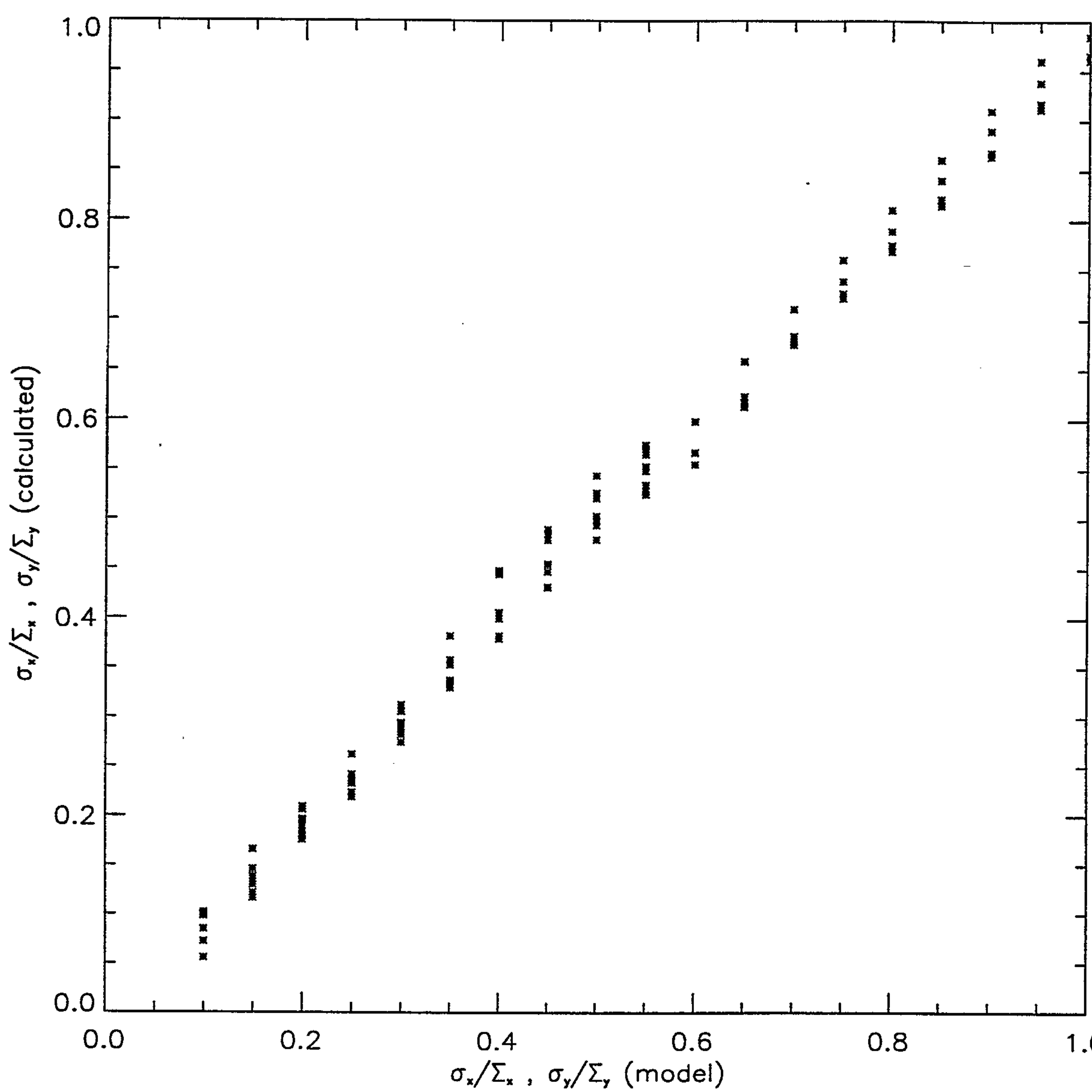

Figure 4: Comparison between original input errors and estimated values obtained from our algorithm, for many different realizations.

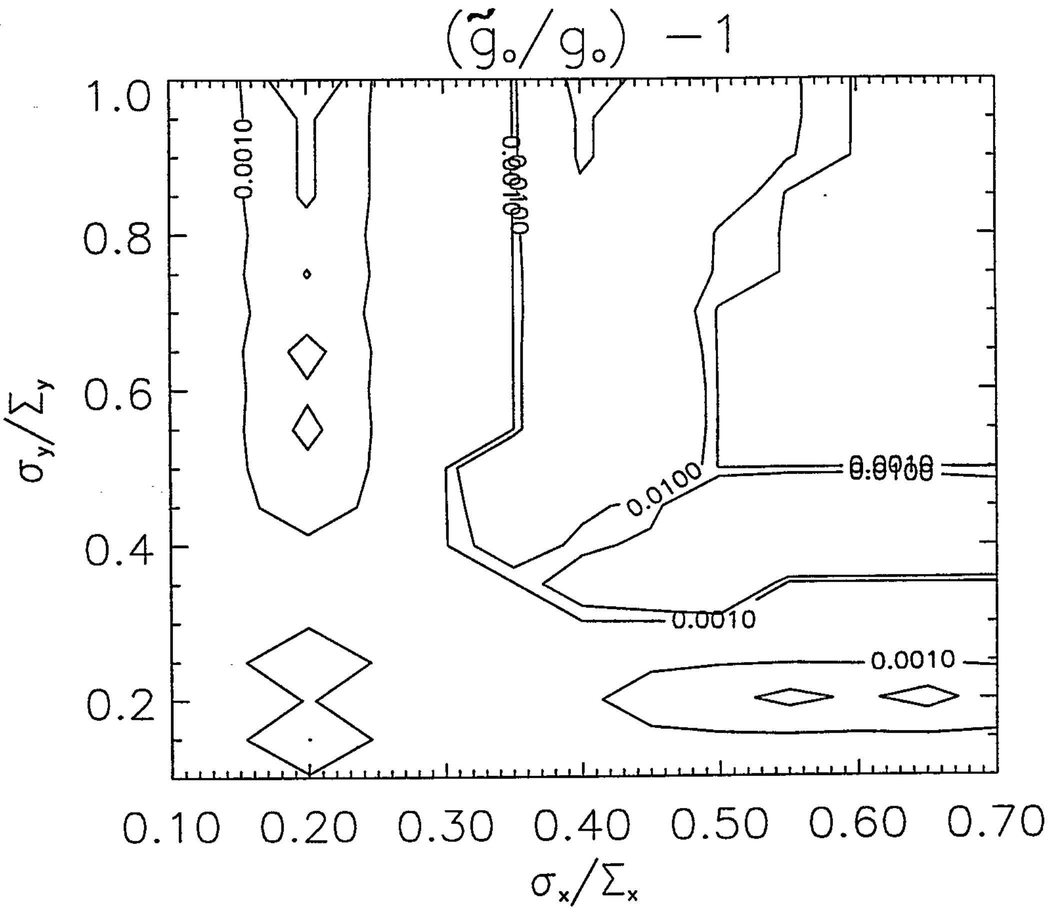

Figure 5: Differences between our estimate of $\tilde{g}_o$ relative to g_o after subtraction of unity.

σ_x \ σ_y	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
4	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	5	5
6		1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3
8			1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3
10				1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3
12					1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
14						1	1	1	1	1	1	1	1	1	1	1	1	1	1
16							1	1	1	1	1	1	1	1	1	1	1	1	1
18								1	1	1	1	1	1	1	1	1	1	1	1
20									1	1	1	1	1	1	1	1	1	1	1
22										1	1	1	1	1	1	1	1	1	1
24											1	1	1	1	1	1	1	1	1
26												1	1	1	1	1	1	1	1
28													1	1	1	1	1	1	1
30														1	1	1	1	1	1
32															1	1	1	1	1
34																1	1	1	1
36																	1	1	1
38																		1	1
40																			1

Table 4: Number of zero-crossings of Ξ for a particular combination of σ_x and σ_y, for $\Sigma_x = 40$ and $\Sigma_y = 40$.

For completeness, we show in Table 4 an example of the number of zero-crossings of Ξ for particular combinations of σ_x and σ_y. Clearly, for a large number of combinations there is only one minimum. However, there are cases, especially for large σ_y and small σ_x, (where we emphasize that these are the "true" intrinsic errors), for which there are several zero-crossings. As already noted, we recommend dealing with cases where the user knows that $\sigma_x < \sigma_y$, and we would not recommend using our technique if the user detects multiple zero-crossings in the region $\tilde{\sigma_x}/\tilde{\Sigma}_x \le \tilde{\sigma_y}/\tilde{\Sigma}_y$. In fact, most of the "extra" zero crossings associated with the cases of small σ_x and large σ_y just discussed (see Table 4) arise in fact when our assumed $\tilde{\sigma_x}$ values exceed $\tilde{\sigma_y}$. This is a nonsense solution and is avoided if we do not carry our iterations in Table 1 past the case $\tilde{\sigma_x} = \tilde{\sigma_y}$.

Nevertheless, There are clearly a wide range of cases in which we can uniquely define estimates of g_o, σ_x, and σ_y. If the user can determine that the error in one coordinate is less than the other, and uses this as the x variable (abscissa), then the multiple crossings are rarely a concern. The first zero encountered is generally the correct combination of $\tilde{g_o}$, $\tilde{\sigma_x}$, and $\tilde{\sigma_y}$.

4 Conclusions

A new method for least-squares fitting of straight-line data has been presented which permits optimal fitting and simultaneous determination of intrinsic errors for both the abscissa and the ordinate. It assumes that the original data and the intrinsic errors are normally distributed. The algorithm works best if the user knows which of the two variables has the smallest intrinsic error, but is not limited to this condition. There are occasions when the procedure has multiple solutions, but these are relatively rare and can be dealt with if the user exercises due care. We anticipate that this process should find application in a wide range of situations.

Acknowledgments

We would like to thank Dr. S. R. Valluri for his review comments and suggestions for improvement of this work.

References

[1] Barker, D. R., Simple method for fitting data when both variables have uncertainties, *Am. J. Phys.,*, 42(3), 224-227, 1974.

[2] Bevington, P. R., Data reduction and error analysis for the physical sciences, *McGraw-Hill,* New York, 1969.

[3] Brandt, S., Statistical and computational methods in data analysis, *Elsevier*, New York, 1976.

[4] Daniel, C. and Wood, F. S., Fitting equations to data, *Wiley-Interscience,* New York, 1971.

[5] Jolivette, P. L., Least-squares fits when there are errors in X, *Computers in Physics,* 7, 208-212, 1993.

[6] Lybanon, M., A better least-squares method when both variables have uncertainties, *Am. J. Phys.,* 52(1), 22-26, 1984.

[7] Miller, B. P. and Dunn, H. E., Orthogonal least-squares line fit with variable scaling, *Computers in Physics,* July/August, 1988.

[8] Orear, J., Least-squares when both variables have uncertainties, *Am. J. Phys.,* 50(10), 912-916, 1982.

[9] Reed, B. C., Linear least-squares fits with errors in both coordinates, *Am. J. Phys.,* 57(7), 642-646, 1989.

[10] Reed, B. C., Linear least-squares fits with errors in both coordinates: comments on parameter variances, *Am. J. Phys.,* 60(1), 59-62, 1992.

[11] Taylor, J. R., An introduction to error analysis, *Mill Valley,* California, U.S.A., 1982.

[12] York, D., Least-squares fitting of a straight line, *Can. J. Phys.,* 44, 1079-1086, 1966.

[13] Young, H. D., Statistical treatment of experimental data, *McGraw-Hill,* New York., 1962.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in section 8.) Defence Research Establishment Ottawa Department of National Defence Ottawa, Ontario Canada K1A 0Z4	**2. SECURITY CLASSIFICATION** (overall security classification of the document including special warning terms if applicable) UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C or U) in parentheses after the title.)

A Least-Squares Straight-line Fitting Algorithm with Automatic Error Determination (U)

4. AUTHORS (Last name, first name, middle initial)

Thayaparan, Tayananthan and Hocking, Wayne, K.

5. DATE OF PUBLICATION (month and year of publication of document) July 1998	**6a. NO. OF PAGES** (total containing information. Include Annexes, Appendices, etc.) 15	**6b. NO. OF REFS** (total cited in document) 13

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.

DREO TECHNICAL NOTE

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address.

Defence Research Establishment Ottawa
Department of National Defence
Ottawa, Ontario Canada K1A 0Z4

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant) 05EE17	**9b. CONTRACT NO.** (if appropriate, the applicable number under which the document was written)
10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.) DREO TECHNICAL NOTE 98-004	**10b. OTHER DOCUMENT NOS.** (Any other numbers which may be assigned this document either by the originator or by the sponsor)

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Distribution limited to defence departments and defence contractors; further distribution only as approved
( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
( ) Distribution limited to government departments and agencies; further distribution only as approved
( ) Distribution limited to defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). however, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

(U) We demonstrate a new generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets which are expected to be linearly correlated are subject to different uncertainties in their measurements. The algorithm determines not only the optimum slope, but also produces estimates of the intrinsic errors associated with each data set. It requires almost no initial information about the errors in each technique.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases thatcharacterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Least-Squares Fitting
Error Analysis
Linear Regression
Data Analysis
Normal distribution
Uncertainties
Measurement Errors
Standard deviation
Variance
Slope
Gradient

#509992