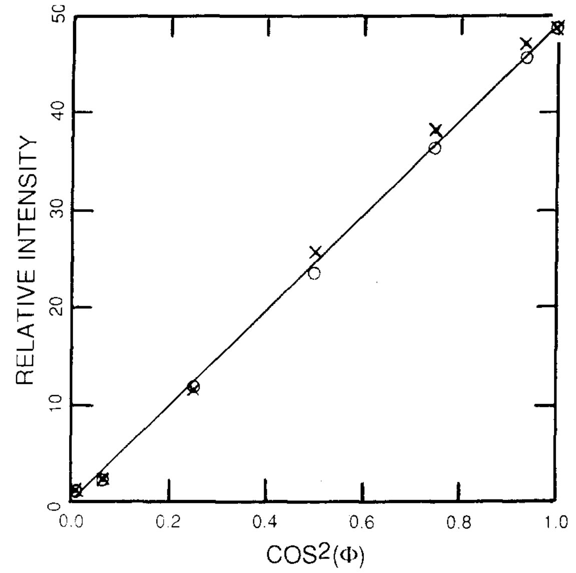

Fig. 6. Transmission of waves through the wire grid polarization analyzer. ϕ is the angle between the grid wires and the horizontal when the E vector of the waves is vertical. The open circles are for clockwise rotation of the grid viewed from behind the receiver; the X dots are for counterclockwise rotation.

output voltage when using the analyzer grid was a function of the distance of the grid from the receiver. To obtain consistent results, it was found necessary to make small adjustments in the grid-to-horn distance at each value of ϕ so as to maximize the receiver output.

With these modifications and procedures the wave intensity at the receiver followed the $\cos^2 \phi$ relationship for analyzer grid rotation, as shown in Fig. 6.

VI. CONCLUSIONS

Our experience in using 3-cm electromagnetic waves shows the necessity of reducing wave reflections from equipment components and surroundings in order to obtain results consistent with theoretical predictions. With the equipment modifications and operating procedures described, students can obtain single slit diffraction patterns for a range of slit widths that cover the transition from Fraunhofer to Fresnel diffraction and also investigate the polarization of the waves.

A listing of the program SLIT may be obtained from the author.

[1] The equipment on which this is based was obtained from Pasco Scientific, 10101 Foothills Blvd., Roseville, CA 95661. Similar equipment may be obtained from Sargent-Welch Scientific Co., 7300 N. Linder Ave., Skokie, IL 60077; from Central Scientific Co., 11222 Melrose Ave., Franklin Park, IL 60131; and other suppliers of educational equipment.

[2] For example: E. Hecht, *Optics* (Addison-Wesley, Reading, MA, 1987), 2nd ed.; G. R. Fowles, *Introduction to Modern Optics* (Holt, Rinehart and Winston, New York, 1975), 2nd ed.; C. L. Andrews, *Optics of the Electromagnetic Spectrum* (Prentice-Hall, Englewood Cliffs, NJ, 1960).

[3] *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1965), pp. 301–302.

Least-squares fitting when both variables contain errors: Pitfalls and possibilities

J. Ross Macdonald and William J. Thompson
Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27599-3255

(Received 8 April 1991; accepted 31 May 1991)

Least-squares fitting is reviewed, in tutorial form, when both variables contain significant errors. Various error models are described; corresponding appropriate weighting is discussed; and the interpretation of weighting is clarified by a physically intuitive description and by graphical results. Resources in the literature on least-squares fitting that are suitable for physics and astronomy students are reviewed. Algorithms for straight-line fitting, indicate practical solution methods, are summarized and numerical comparisons are given. Also described are several readily available computer programs that allow fitting for both straight-line and nonlinear situations and that are appropriate for both research and teaching applications.

I. INTRODUCTION

Least-squares fitting when both variables have errors is a perennially interesting problem, on which a dozen communications have been published in this Journal in the past 2 decades.[1-12] There is also an extensive research literature on the subject, dating back more than a century. The problem is often called generalized least squares in the physical sciences and the errors-in-variables (EOV) model in statistics.[13] Such least-squares fitting is also of significant interest in astronomy[14] and in chemistry.[15] For students in the physical sciences it is important to learn the pitfalls and possibilities in least-squares analyses, especially since many of the algorithms are now available in user-friendly programs whose underlying models may not be understood by the user.

The major purposes of the present paper are to review and clarify for teachers and students the common points and differences in previous presentations, to systematize current understanding of least-squares fitting when both variables contain errors, and to list some relevant and available computer programs. In spite of previous consideration, new information is available, problems previously unmentioned need attention, and future directions for investigation are indicated.

Since confusion may arise from the varying meaning of "linear" by different authors,[9,11,12] some clarification is worthwhile. "Linear least squares" properly means that the parameters of the fitting model appear only linearly in it, no matter whether the model is linear or nonlinear in its variables, e.g., $y = a + bx + cx^2$ is a linear model in terms of parameters a,b,c. When some authors refer to "linear least squares," however, they mean linear in the variables, i.e., $y = a + bx$, which we will call "straight-line least squares." Thus "straight-line least-squares fitting" and "linear least-squares fitting" are not necessarily synonymous. Many of the papers in this journal deal only with fitting a straight line, but some treat nonlinear least-squares situations, where the functional relations between data and parameters are nonlinear, and a few of the papers describe models that are general enough to accommodate more than two variable types.

The outline of this paper is as follows. We discuss in Sec. II A how the distributions of errors in the variables can be handled, and in Sec. II B how weights may be assigned to the data in forming the least-squares objective function. The common special case of straight-line least-squares fitting is discussed in Sec. III, where numerical comparisons are also presented. Future directions are indicated in Sec. IV, and a brief discussion of currently available computer programs for general EOV fitting is given in Sec. V.

II. ERROR AND WEIGHTING MODELS FOR LEAST-SQUARES FITTING

In least-squares fitting models there are two distinct ingredients that are usually discussed in only a cursory way; the first is a model for the errors in the variables, and the second is a model for the weight that each datum has in the fitting. We now summarize relevant concepts and techniques related to these submodels, using straight-line fitting if an example is appropriate.

A. Error models

When both variables contain errors, any distinction between dependent and independent variables is ambiguous, although one usually attempts to control one of them, x, and observe the other, y. In order to characterize the problem, assume that single measured values have the form $x_i = X_{0i} + \epsilon_{xi}$ and $y_i = Y_{0i} + \epsilon_{yi}$ for each of the $i = 1,2,...,N$ data pairs. The exact values X_{0i} and Y_{0i} satisfy a functional relationship, for example, $Y_{0i} = a_0 + b_0 X_{0i}$ in the straight-line case. But the errors, ϵ_{xi} and ϵ_{yi}, are unknown, except in Monte Carlo simulations. One needs to make assumptions about these unknown errors, in order to pick an appropriate solution method. We also define fitted values, X_i and Y_i, related in the straight-line case by $Y_i = a + bX_i$. The major objective of least-squares fitting is to obtain estimates of X_i and Y_i as close to X_{0i} and Y_{0i} as possible, so that for straight-line fits a and b are minimum-bias estimates of a_0 and b_0.

In ordinary least-squares analyses, two different error assumptions are common, leading to[16] the *homoscedastic standard error model* and to the *heteroscedastic diagonal error model*. The two models may be characterized by the following expectation values (E in statistical notation). For both models, the average error in x and y is zero at each datum: $E(\epsilon_{xi}) = 0$ and $E(\epsilon_{yi}) = 0$ (systematic errors are absent), and from point to point the errors are uncorrelated; thus $E(\epsilon_{xi}\epsilon_{xj}) = 0$ and $E(\epsilon_{yi}\epsilon_{yj}) = 0$ for $i \neq j$. For homoscedastic errors, the variances are assumed to be independent of the data point: $E(\epsilon_{xi}^2) = \sigma_x^2$ and $E(\epsilon_{yi}^2) = \sigma_y^2$. For heteroscedastic errors, the errors are heterogeneous from point to point: $E(\epsilon_{xi}^2) = \sigma_{xi}^2$ and $E(\epsilon_{yi}^2) = \sigma_{yi}^2$. We have not specified the error covariances, $E(\epsilon_{xi},\epsilon_{yj})$, but they are usually assumed to be zero (uncorrelated x and y errors). A restricted form of this model is $\sigma_{yi}^2/\sigma_{xi}^2 \equiv \lambda_0$, a constant *ratio* of error variances for all data. When $\lambda_0 = \infty$, there are errors only in y, while for $\lambda_0 = 0$ the errors are only in x.

If the x and y errors are uncorrelated from point to point, then x or y error values are random samples from independent probability distributions with zero mean, which we write as $\epsilon_{xi} = \sigma_{xi} P_x(0,I_i)$ and $\epsilon_{yi} = \sigma_{yi} P_y(0,I_i)$, where P_x and P_y are independent distributions and $I_i \equiv 1$ is an element of the unit vector of dimension N. For completeness, the *general error model* is like the diagonal error model but has $E(\epsilon_{xi}\epsilon_{xj}) = \sigma_{xij}^2$ and $E(\epsilon_{yi}\epsilon_{yj}) = \sigma_{yij}^2$ for various i and j values. In weighted least squares, one of the biggest problems is to obtain appropriate estimates of the error variances, and the problem is greatly exacerbated when error covariances such as σ_{yij}^2 $(i \neq j)$ must also be estimated. Thus there are many formulations of this model but very few applications of it.

B. Weighting models

For the weighting model (called a variance model in statistics) that takes into account the relative influence of each datum in a least-squares fit, one needs to capture the effects of the unknown errors and to compensate for them in optimum fashion. As Deming[17] proposed, an appropriate way to do this is to use weighted least squares and to find those sets of X_i and Y_i which minimize the objective function

$$O = \sum_{i=1}^{N} [w_{xi}(x_i - X_i)^2 + w_{yi}(y_i - Y_i)^2], \qquad (1)$$

where the weights, w_{xi} and w_{yi}, are often the reciprocals of the estimated variances for the observations. The weights should certainly be dimensionally consistent with the x and y variables so that O is dimension free. Physicists usually refer to O as chi-squared, but if other prescriptions for the weights are used, then O is unlikely to be distributed as the statisticians' chi-squared. In fact, the above expression, even with the proper weights, is not exactly so distributed when both x and y weights are nonzero.[18] It is implicit in Deming and in most succeeding work that although the estimated weights may vary with i, they are independent of the calculated fitting variable values, X_i and Y_i, for all values of i and can thus be taken constant when minimizing O with respect to the fitting parameters. But, as discussed in

Table I. Acronyms for weighting models for least-squares fitting with errors in both variables, as discussed in Sec. II B. The acronyms are ordered from most general to least general.

Acronym	Weighting model	Summary of definition
	Off-diagonal weights	
IGWM	Independent general weighting model	w_{xi} and w_{yi} independent
	No off-diagonal weights	
DWM	Dependent weighting model	w_{xi} and w_{yi} dependent
PWM	Proportional weighting model	$s_{xi}/X_i = \text{(const)}$ $s_{yi}/Y_i = \text{(const)}'$
IDWM	Independent diagonal weighting model	w_{xi} and w_{yi} independent
IDWMC	Independent diagonal weighting model with constant weight ratio	$w_{xi}/w_{yi} = \lambda$
SWM	Standard weighting model	$w_{xi} = w_x,\ w_{yi} = w_y$ $w_x/w_y = \lambda$
OLS-*y*:*x*	Ordinary least squares—*y* on *x* with *y* errors only	$s_{xi} = 0$
OLS-*x*:*y*	Ordinary least squares—*y* on *x* with *x* errors only	$s_{yi} = 0$

Sec. IV, in many situations such constancy is unjustified and use of inappropriate fixed weights can lead to inaccurate parameter estimates, that is, to parameter bias.

Thus one needs to further distinguish between fitting models in which the errors and corresponding weights are independent of the fitting variables—the *independent diagonal weighting model* (IDWM), and those in which they are *dependent* (DWM). Finally, the *independent general weighting model* (IGWM) includes off-diagonal as well as diagonal terms in the weighting matrix. Table I lists and defines the weighting acronyms used in the present work.

If the estimated standard deviations of individual observations are denoted as $s_{xi} \simeq \sigma_{xi}$ and $s_{yi} \simeq \sigma_{yi}$, then the corresponding weights are $w_{xi} = 1/s_{xi}^2$ and $w_{yi} = 1/s_{yi}^2$. For the IDWM, the weighting model is just defined by the estimates of w_{xi} and w_{yi}. When their ratio is *constant* (the IDWMC), we may define $w_{xi}/w_{yi} = s_{yi}^2/s_{xi}^2 \equiv \lambda$, an estimate of λ_0. Of course when the weights themselves are independent of *i*, then $w_{xi}/w_{yi} = s_y^2/s_x^2 \equiv \lambda$, called the standard weighting model, SWM.

If there are constant proportional errors in the *x* data and constant proportional errors in the *y* data, then one should use the *proportional weighting model*, PWM, a particular type of DWM. Since the errors are proportional to the error-free, unknown quantities X_{0i} and Y_{0i}, the PWM involves s_{xi} proportional to X_i and s_{yi} proportional to Y_i, with X_i and Y_i taken as the best available estimates (after iterative convergence) of X_{0i} and Y_{0i}, respectively. If the weighting standard deviations were instead chosen proportional to x_{0i} and y_{0i}, bias would be introduced in the parameter estimates since the data values already include errors. Unless $Y \propto X$, the PWM does not lead to a constant λ. With the nomenclature introduced above, one can characterize the weighting models of various authors, as shown in Table II.

Mechanical analogies, illustrated in Fig. 1 for six different weighting schemes, help clarify the distinction between the various models. In Eq. (1) if the weights were literally weights, then the first term in the sum would be the moment of inertia of a distribution of mass points *i* about the X_i, and the second term would be the moment of another distribution about the Y_i. Minimizing *O* therefore minimizes the sum of these two moments of inertia. In the IDWMC, the two mass distributions differ only by an overall scaling factor λ so they are essentially the same and one is minimizing the total moment of inertia about an axis which is the best-fit straight line. For the SWM the weights are all equal, so the moments about this axis will differ from those of an equal-mass-point system only if the spacing of data points is nonuniform. A somewhat similar geometric representation is discussed in Lybanon.[9]

III. STRAIGHT-LINE FITTING

Now that we have summarized the statistical terminology in Sec. II, we are ready to discuss various methods for straight-line fitting. In the following we survey appropriate

Table II. Synopsis of least-squares analysis models with errors in both variables. The author list is chronological by year of first reference cited. The weighting models are discussed in Sec. II B, and Fig. 1 illustrates IDWM and SWM.

Authors; year (Ref.)	Weighting model	Approximate or exact	Non-linear, NL linear, L or straight line, SL
Deming; 43,64[17]	IDWM	A	NL
Madansky; 59[32]	SWM	E	SL
York; 66[24]	IDWM	E	SL
Gerhold; 69[1]	IDWM	E	L
Powell *et al.*; 72[27]	IDWM	E	NL
Britt *et al.*; 73–75[3,28,33]	IGWM	E	NL
Barker *et al.*; 74[2]	IDWM	A	L
Macdonald; 75[4]	IDWM	A	SL
Riggs *et al.*; 78[20]	SWM	E	SL
Jefferys *et al.*; 80–90[21,22]	IGWM	E	NL
Ross; 80[5]	SWM	A	SL
Krane *et al.*; 82[7]	IDWM	E	SL
Orear; 82[8]	SWM	A	NL
Lybanon; 84[9,10]	IDWM	E	L
Christian *et al.*; 84–86[15,23]	IDWM	A	NL
Fuller; 87[13]	IDWM	E	NL
Reed; 89[11]	IDWM	E	SL
Squire; 90[12]	IDWM	A	NL

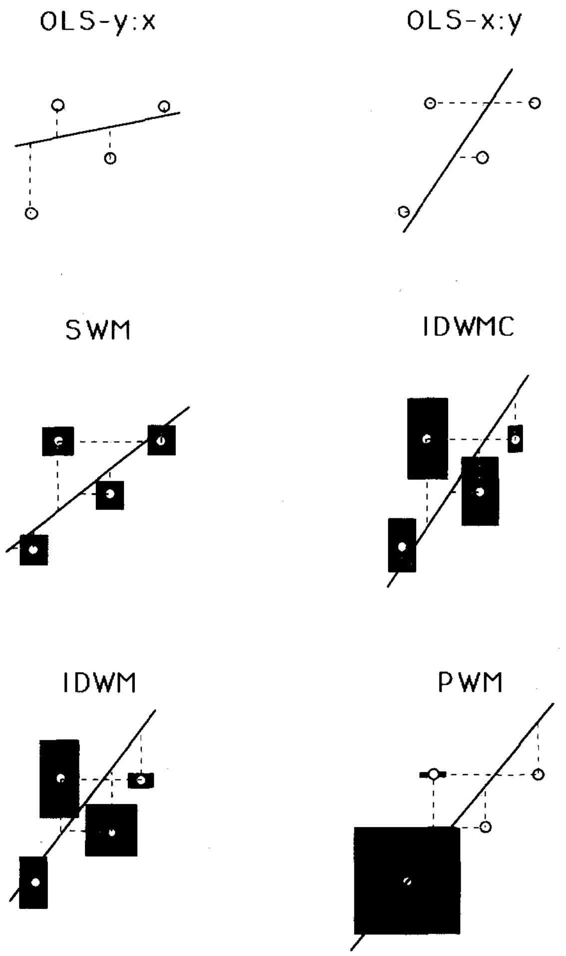

Fig. 1. Weighting models illustrated for straight-line least-squares fits. The data values, $(x,y) = (1,1),(2,5),(4,3)$, and $(6,5)$, are centered in the circles and are the same for all fits. Shaded rectangles indicate weights, with x weights, w_{xi}, as widths in the x direction and y weights, w_{yi}, as y heights, if they are not smaller than the circles. Dashed lines show differences whose weighted sum-of-squares is minimized. Solid lines are least-squares fits obtained with the program GENLS. In OLS-y:x one minimizes unweighted y differences; in OLS-x:y unweighted x differences are minimized; in SWM all x and y weights are equal (square weights); in IDWMC the ratio of weights λ is constant (weights of constant shape); in IDWM weights vary independently; and in PWM the weights are inversely proportional to the squares of the X and Y values.

algorithms, some simplifications of these, and we present numerical comparisons.

A. Survey of algorithms

Many scientific data-analysis problems can be analyzed by straight-line fitting after appropriate transformation of variables, albeit often with introduction of bias in parameter estimates.[19] Further, the complexity of the analysis is greatly reduced for straight-line fits. Therefore, this survey emphasizes straight-line fitting algorithms.

Table II summarizes representative papers of the past 40 years, particularly those appearing in this Journal. Not widely known are the extensive Monte Carlo simulation studies carried out by Riggs *et al.*,[20] who discussed 34 different straight-line solutions and investigated their applicability for the standard error model using the SWM. There is an extensive literature on least squares in astronomy,[14,21,22] an observational science in which the distinction between dependent and independent variables is usually irrelevant. The pedagogical literature in chemistry (such as Refs. 15 and 23) also discusses straight-line algorithms.

An important advance in straight-line fitting for the IDWM (including the SWM) was made by York.[24] It has recently been discussed by Reed,[11] who points out that it is exact (see Refs. 2, 4, and 23 for earlier recognition of this), and who discuss methods of calculating the slope estimate b from York's quasicubic equation. Earlier calculational approaches for the IDWM were described by Gerhold[1] and Williamson[25] (for straight-line situations only). Although Williamson's iterative solution is based on that of York, it requires no root selection. Squire[13] responded to Reed's work by emphasizing that any solution requiring the selection of roots, such as that of York and Reed, cannot be strongly recommended. Instead, he suggested using a more general method allowing solution of the IDWM for nonlinear situations as well. The computer program that he recommended, NLLSQ, is not currently available.[26] Further, the method that Squire suggested yields an exact solution only for straight-line fits.

Two general algorithms without such defects had already been devised by Powell and Macdonald,[27] and by Britt and Luecke.[28] They were the first algorithms for nonlinear situations that converged to exact solutions upon iteration. The Britt and Luecke method is more general than that of Powell and Macdonald, but it requires analytic expressions for first derivatives. The Powell–Macdonald approach uses special numerical derivatives that greatly speed convergence, and it has recently been generalized and incorporated in the program GENLS discussed in Sec. V.

Later, Jefferys[21] described an algorithm very similar to that of Britt and Luecke but generalized to allow additional constraint equations, which are often needed in astronomy. It was later embodied in the program GaussFit.[22] When applied to the same data and fitting model, the Jefferys approach leads[9] to the same converged solution as does the Powell–Macdonald algorithm,[27] suggesting that they both properly yield least-squares results.

B. Simplifications and practical methods

As Squire[12] pointed out, a general program is valuable for solving the IDWM problem for either linear or nonlinear situations. Such a program should also yield the York solution for straight-line fits. The algorithms described in Refs. 22, 27, and 28 do so, and they avoid the need to choose one of three roots to estimate the slope. For those who do not have the appropriate program (Sec. V), it is worthwhile to consider useful simplifications of York's results that follow[24] when the ratio of the weights is independent of i, that is, the IDWMC in Fig. 1.

First, form the total weight for the ith data point W_i that associated with the effective variance approach to straight-line EOV fitting[2,24,29-31], in terms of the slope b,

$$W_i = (w_{yi}^{-1} + b^2 w_{xi}^{-1})^{-1}. \tag{2}$$

Now for the IDWMC, we can express W_i as

$$W_i = w_{xi}(\lambda + b^2)^{-1} = w_{yi}(1 + b^2/\lambda)^{-1}. \tag{3}$$

Define the weighted sums over i from 1 to N: $S_w \equiv \Sigma W_i$; $S_x \equiv \Sigma W_i x_i$; and $S_y \equiv \Sigma W_i y_i$. Then the weighted means of x and y are

$$\bar{x} = S_x/S_w \tag{4}$$

and

$$\bar{y} = S_y/S_w. \tag{5}$$

We may now define

$$S_{xx} \equiv \sum W_i (x_i - \bar{x})^2, \tag{6}$$

$$S_{yy} \equiv \sum W_i (y_i - \bar{y})^2, \tag{7}$$

and

$$S_{xy} \equiv \sum W_i (x_i - \bar{x})(y_i - \bar{y}), \tag{8}$$

which are analogous to weighted variances and covariances.

If λ is a constant, then so is

$$\gamma \equiv (\lambda S_{xx} - S_{yy})/2S_{xy}. \tag{9}$$

York's pseudocubic then reduces[24] to the quadratic $(b^2 + 2\gamma b - \lambda) = 0$. Although this might appear to be a quasiquadratic because the W_i implicit in γ involve b^2, as shown in Eq. (3), this is not so. Since the denominators in Eq. 3) are independent of i, they cancel in Eq. (9), and we may replace the W_i in Eq. (9) by either w_{xi} (for $0 \leqslant \lambda < \infty$) or by w_{yi} (for $0 < \lambda \leqslant \infty$). When such replacement is made in Eq. (9), solution of the quadratic (which is maximum likelihood when the errors are normally distributed), gives[24,32] for the dimensionless quantity $\beta \equiv b/\sqrt{\lambda}$:

$$\beta = -(\gamma/\sqrt{\lambda}) + \mathrm{sgn}(S_{xy})\sqrt{1 + (\gamma^2/\lambda)}$$
$$= 1/\left[(\gamma/\sqrt{\lambda}) + \mathrm{sgn}(S_{xy})\sqrt{1 + (\gamma^2/\lambda)}\right], \tag{10}$$

where b is the estimated slope and $\mathrm{sgn}(S_{xy})$ selects the appropriate root. Numerical accuracy is best with the first form when γ and $\mathrm{sgn}(S_{xy})$ have opposite signs, but with the second for the same signs.

When γ involves only the w_{xi} weights, Eq. (10) is equivalent to that given by Deming.[17] Although Deming applied his result whether or not λ was a constant, only when λ is constant does Eq. (10), or the more general Deming algorithm,[17] yield an exact least-squares solution.[24,33] Although the slope estimate b obtained from Eq. (10) would be minimum bias for the IDWMC if $w_{xi} = 1/\sigma_{xi}^2$, $w_{yi} = 1/\sigma_{yi}^2$, and λ_0 were used in calculating γ, in practice one must use the estimates $w_{xi} = 1/s_{xi}^2$, $w_{yi} = 1/s_{yi}^2$, and λ, with consequent introduction of unknown bias. A promising alternative is outlined in Sec. IV.

For the SWM (Fig. 1), all weights cancel in Eq. (10) and only λ and the unweighted sums remain. The resulting expression for b has been named the PW solution (for *perpendicular* least squares, properly *weighted*) by Riggs *et al.*[20] The PW is quite distinct from the PWM discussed in Sec. II B. The PW yields the exact solution when the SWM is appropriate and $\lambda = \lambda_0$. Even for this situation, Riggs *et al.* found that the Eq. (10) estimate b_{PW} is appreciably biased for data with a correlation coefficient less than about 0.7. Thus even the exact solution of the least-squares equations can be biased. Further, they found that b_{PW} is extremely biased whenever λ is a poor approximation to λ_0. Fuller[13] has also discussed such bias effects.

In all the above analyses, whenever a slope estimate has been obtained, the corresponding estimate of the intercept, a, follows from

$$a = \bar{y} - b\bar{x}. \tag{11}$$

Therefore, bias in b will necessarily cause bias in a.

Further simplifications are worth mentioning. When $\lambda \to \infty$, so that $w_{xi} \to \infty$ and $W_i = w_{yi}$, Eq. (10) leads to an estimate of the slope of y on x, for relatively negligible x errors,

$$b_{y:x} = S_{xy}/S_{xx}. \tag{12}$$

This is an *ordinary least squares* fit of y as a function of x, which we show as OLS-*y*:*x* in Fig. 1. Alternatively, when $\lambda \to 0$, so that $w_{yi} \to \infty$ and W_i may be replaced by w_{xi}, one obtains for relatively negligible y errors,

$$b_{x:y} = S_{yy}/S_{xy}. \tag{13}$$

This may be termed OLS-*x*:*y* fitting, as shown in Fig. 1, but it is expressed in terms of the *y*-on-*x* slope, to allow direct comparison with the Eq. (12) result.

Finally, consider the SWM situation and use the estimated standard deviations of the data for weighting. Thus take

$$s_{xi}^2 \equiv \frac{1}{w_{xi}} = (N-1)^{-1} \sum (x_i - \bar{x})^2 \equiv s_x^2,$$

and

$$s_{yi}^2 \equiv \frac{1}{w_{yi}} = (N-1)^{-1} \sum (y_i - \bar{y})^2 \equiv s_y^2,$$

for all i. Then Eq. (9) yields $\gamma = 0$ and $b = b_{\mathrm{GM}}$, where

$$b_{\mathrm{GM}} = \mathrm{sgn}(S_{xy})(s_y/s_x) = \mathrm{sgn}(S_{xy})\sqrt{S_{yy}/S_{xx}}$$
$$= \mathrm{sgn}(S_{xy})\sqrt{b_{y:x} b_{x:y}}. \tag{14}$$

This estimate is the *geometric mean* (GM) of the two ordinary regression slopes and thus must lie between them, as does the slope following from the PW solution. Clearly, one can obtain b_{GM} by making two fits of the data with an ordinary linear least-squares fitting program, then using Eq. (14). Riggs *et al.*[20] consider this result to be of central importance, partly because it appears to be independent of λ. Since, with the choice of weights just made, b_{GM} can be expressed as $\mathrm{sgn}(S_{xy})\sqrt{\lambda}$, it seems improper to consider λ and S_{yy}/S_{xx} as separate variables in GM fitting. The GM solution thus implies that

$$\lambda = \lambda_{\mathrm{GM}} \equiv S_{yy}/S_{xx} \equiv b_{\mathrm{GM}}^2, \tag{15}$$

entirely determined by the data and not available for adjustment.

C. Numerical comparisons

Now that we have a variety of models, it is interesting to compare them when applied to representative data. Before launching into the details of numerical comparisons of straight-line least-squares algorithms, we point out a physics-wise and pedagogically inappropriate aspect of earlier analyses. Namely, the dimension of the parameter λ has been systematically ignored, in spite of the fact that its dimension is that of the square of y/x. The parameter β in Eq. (10) is an appropriate dimension-free parameter. In the examples discussed below, *x*,*y*, and therefore λ, are implicitly dimension free.

We first consider analysis of the data in Reed's[11] Table I, where he set $s_x = s_y = 0.01$, so $\lambda = 1$, a SWM situation. When the PW is used, one immediately finds his solution: $Y \simeq -0.365 + 1.167X$. But in obtaining his direct solution of the York pseudocubic he spent 30 iterations that did not converge to a root, and 9 more iterations from a different starting value that did converge. Clearly, as York implied, when a constant λ is assumed it is far more efficient to use the PW simplification, Eqs. (10) and (11), for the IDWM or the SWM.

But what about the even simpler GM solution? For Reed's Table I data it yields $Y \simeq -0.336 + 1.135X$, not far from the PW result. If the intercept estimate value is calculated as the average of the y intercepts for the OLS y:x and x:y solutions instead of from Eq. (11), one obtains a $\simeq -0.357$, somewhat closer to the PW value. But, according to Eq. (15), $\lambda_{GM} \simeq 1.288$, appreciably different from the value of unity assumed in Reed's PW solution.[11] If one believed that the $\lambda = 1$ choice (in dimension-free units) were accurate, then one could immediately calculate the GM slope as $\sqrt{\lambda}$, obtaining unity here, somewhat farther from the PW solution than that calculated from Eq. (14). In their Monte Carlo study Riggs *et al.*[20] found that the Eq. (14) solution was essentially unbiased for $\lambda_0 = 1$, but generally one does not know λ_0.

Finally, it is worthwhile to examine an SWM situation where λ is not expected to be close to unity. We select the 15-point data in Table 1.3.1 of Fuller,[13] who took $\lambda = 1/6$. For this λ the PW yields $Y \simeq (1.1158 \pm 0.89) + (0.7516 \pm 0.087)X$. The $\pm$ terms are estimated standard deviations[17,24] and the parameter estimates are shown with more precision than is justified by their uncertainties to permit comparison with Fuller's results. The parameter estimates agree fully with his values. The GM leads to $Y \simeq (1.4867 \pm 0.87) + (0.7147 \pm 0.085)X$, statistically consistent with the PW results. Further, $\lambda_{GM} \simeq 0.511$, but the square root of the assumed value of 1/6 is about 0.408, not a good estimate of the likely b value, thus possibly indicating a poor weighting choice. If we follow Riggs *et al.*[20] and define $K^2 \equiv \lambda / b^2_{GM} = \beta^{-2}_{GM}$, where λ is the assumed value, then K^2 is about 0.326. For $\lambda = 1$, the PW result becomes $Y \simeq (1.715 \pm 0.85) + (0.692 \pm 0.084)X$, $K^2 \simeq 1.96$ and the GM solution remains the same. If we select $\lambda = b^2_{GM} \simeq 0.5107$, so that $K^2 = 1$, then the PW and the GM solutions are identical.

The above numerical comparisons suggest that whenever a constant ratio of errors is believed to be appropriate, it is worthwhile to calculate both the PW and GM solutions and to compare them. To minimize bias in the estimated parameter values, as good an estimate as possible for λ is needed. When λ is very uncertain, the GM solution is likely to be better than the PW solution.

IV. FUTURE DIRECTIONS

The purpose of least-squares fitting is to obtain accurate estimates of parameters with uncertainties as small as possible. The Monte Carlo results of Riggs *et al.*[20] show that, even with true minimization of Eq. (1) for straight-line fitting using the SWM and a solution such as York's, the parameter biases are usually nonzero (instead of zero as in conventional straight-line least squares fitting), and they are usually functions of the weighting-ratio parameter λ. Even if s_x^2 and s_y^2 were accurate estimators of σ_x^2 and σ_y^2, a poor solution might still be obtained. For experimental data, estimates of the s_{xi}^2 and s_{yi}^2 values used for weighting are often quite far from optimum, leading to even worse results. Under such conditions, it scarcely seems worthwhile to expend appreciable effort on obtaining a complicated exact solution, and simpler solutions, such as those in Sec. III, may suffice.

Since neither the SWM nor the IDWM is broad enough to represent adequately many types of possible experimental errors, there is a need for programs that include general weighting possibilities for the many-variable situation where the fitting function need not be expressed explicitly in terms of one of the variables, and where nonlinear as well as linear models can be used. We are currently developing methods of treating generalized weighting models dependent on calculated variable values. There are technical difficulties in the general nonlinear EOV situation to estimate simultaneously by maximum likelihood the parameters of both a fitting model and a weighting variance model if the latter may depend on the dependent and independent variables, as in the PWM. Much progress has already been made, however, for complex and real nonlinear least-squares fitting with the independent variable error free.[34,35]

In practice, one would first make a fit with all parameters of both the fitting and the weighting model free to vary. Once meaningful estimates of the weighting-model parameters were obtained, they would be fixed at their estimated values and a final-fit solution obtained. Such an approach would remove some uncertainty in picking the proper weighting; it would reduce bias in parameter estimates; and it would probably also lead to better estimates of the parameter standard deviations. Until the development of an EOV algorithm which allows free parameters in both the fitting and the weighting models, one can use one of the programs discussed below, GENLS, which allows a weighting model dependent on the calculated variable values but with fixed weighting-model parameters.

V. AVAILABLE COMPUTER PROGRAMS

Although the simple PW and GM solutions discussed in Sec. III are usually adequate for straight-line fitting with either the SWM or the IDWM, especially when a good estimate of λ_0 is available, more-general situations often occur. Even for the straight-line fit, the weights may not yield a constant value of λ. The λ value is variable, for example, for the important case of proportional errors, that is, constant percentage errors for the x variable and constant percentage errors for the y variable. GENLS fitting results using the corresponding weighting model, the PWM, are shown in Fig. 1. Further, the fitting model is often nonlinear in its parameters, so more powerful fitting programs are frequently necessary. The following more or less complete programs may be used for both linear and nonlinear situations.

Britt–Luecke: Two implementations of the original Britt–Luecke algorithm are available. The first, GENLSQ, is an IDWM realization of the original IGWM algorithm.[28] It may be obtained from Dr. R. L. Luecke, Department of Chemical Engineering, University of Missouri, Columbia, MO 65211, by sending him a diskette. Only FORTRAN source code is provided; the user must provide code for the fitting function. A commercial version, incorporated in a chemical process simulation system, may be purchased from Dr. H. I. Britt, Aspen Technology, 251 Vassar St., Cambridge, MA. The following two sets of programs involve many elements of the Britt–Luecke approach.

Fuller: Several complete EOV programs are available for IBM PC/AT computers and their clones from the Statistical Laboratory of Iowa State University, Ames, Iowa 50011. Some description of these algorithms appears in Ref. 13. The cost of the programs ranges from $150 to $500. In most cases, they are likely to yield somewhat smaller parameter standard deviation estimates than does GENLS.

GaussFit: The program GaussFit[22] allows fitting with constraint equations. It may be obtained through E-mail by anonymous file transfer program (ftp) to bessel.as.utexas.edu. The GaussFit, Fuller, or Britt–Luecke approaches must be used if the fitting model cannot be expressed as $Y=f(X)$ or $X=g(Y)$, or if more than two kinds of variables are observed. The required first derivatives are automatically calculated symbolically in GaussFit. In addition, it allows the user to define the fitting function, followed by an automatic interpreter step to produce the final program. GaussFit is available for UNIX operating systems and for Macintosh computers; it requires various uncompression procedures and Microsoft Word to obtain the extensive manual; and it also requires a C-language compiler.

GENLS: If the fitting model can be expressed as $Y=f(X)$ or $X=g(Y)$ and no extra constraints are necessary, the GENLS program has many advantages. It uses the Newton–Raphson approach, whereas the Britt–Luecke, Fuller, and GaussFit programs use a Gauss–Newton algorithm. GENLS is available from J. R. Macdonald by sending a formatted disk and a stamped, self-addressed disk mailer, or through E-mail by anonymous file transfer to ftp.oit.unc.edu, directory pub/fitit. It is fully self contained; it includes both FORTRAN source code and a ready-to-use executable program file for MS-DOS operating systems; and it does not require a manual. GENLS provides many different fitting-function choices which can be used without recompilation, but it requires recompilation for fitting functions not included in it, or for use on a Macintosh computer or a machine running UNIX.

GENLS does not require the user to provide analytical expressions for derivatives, and its use of nonconventional numerical derivatives greatly accelerates its approach to convergence.[27] Because it requires the inversion of only a $p \times p$ matrix, where p is the number of free fitting parameters, rather than the generally far larger $N \times N$ matrix of Gauss–Newton algorithms, its execution time is far less for the same problem, particularly when N, the number of data values, is large. GENLS can provide PW and GM solutions for straight-line fitting, as well as the general solution for linear or nonlinear situations. It is fully interactive and menu driven, thus making it very easy in a single run to carry out many different operations, such as changing weightings, observing and saving the results, displaying screen plots of residuals, and varying the number of fitting parameters.

In addition to SWM, IDWMC, and IDWM weighting, GENLS includes a DWM, an important option not available in any other EOV program. To do so, it incorporates a special iterative procedure that allows accurate fitting with combined constant and power-law weighting.[18,19,34,35] It thus includes both PWM and Poisson weighting (appropriate for errors proportional to the square roots of error-free values). Further, different weighting models may be used for the x and for the y parts of the fit, as dictated by the physical situation involved. A future version of GENLS will include more-general weighting possibilities.

[1] G. N. Gerhold, "Least-squares adjustment of weighted data to a general linear equation," Am. J. Phys. **37**, 156–161 (1969).

[2] D. R. Barker and L. M. Diana, "Simple method for fitting data when both variables have uncertainties," Am. J. Phys. **42**, 224–227 (1974).

[3] R. H. Luecke and H. I. Britt, "Parameter estimation with error in the observables," Am. J. Phys. **43**, 372 (1975).

[4] J. R. Macdonald, "Comment on 'Simple method for fitting data when both variables have uncertainties,'" Am. J. Phys. **43**, 372–374 (1975).

[5] A. W. Ross, "Regression line analysis," Am. J. Phys. **48**, 409 (1980).

[6] J. M. Pasachoff, "Applicability of least-squares formula," Am. J. Phys. **48**, 800 (1980).

[7] K. S. Krane and L. Schecter, "Regression line analysis," Am. J. Phys. **50**, 82–84 (1982).

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

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

[10] M. Lybanon, "Comment on 'Least squares when both variables have uncertainties,'" Am. J. Phys. **52**, 276–278 (1984).

[11] B. C. Reed, "Linear least-squares fits with errors in both coordinates," Am. J. Phys. **57**, 642–646 (1989); **58**, 189 (1990).

[12] P. T. Squire, "Comment on 'Linear least-squares fits with errors in both coordinates,'" Am. J. Phys. **58**, 1209 (1990).

[13] W. A. Fuller, *Measurement Error Models* (Wiley, New York, 1987).

[14] *Errors, Bias and Uncertainties in Astronomy*, edited by C. Jaschek and F. Murtagh (Cambridge U. P., Cambridge, 1990) and references therein. See also, T. Isobe, E. D. Feigelson, M. G. Akritas, and G. J. Babu, "Linear regression in astronomy. I." Astron. J. **364**, 104–113 (1990).

[15] S. D. Christian, E. E. Tucker, and E. Enwall, "Least-squares analysis: A primer," Am. Lab. **18**, 41–49 (1986) and references therein.

[16] R. W. Farebrother, *Linear Least Squares Computations* (Dekker, New York, 1988).

[17] W. E. Deming, *Statistical Adjustment of Data* [Wiley, New York, 1943; Dover (reprint), New York, 1964].

[18] J. R. Macdonald, "Review of some experimental and analytical equations of state," Rev. Mod. Phys. **41**, 316–349 (1969).

[19] W. J. Thompson and J. R. Macdonald, "Correcting parameter bias caused by taking logs of experimental data," Am. J. Phys. **59**, 854–856 (1991).

[20] D. S. Riggs, J. A. Guarnieri, and S. Addelman, "Fitting straight lines when both variables are subject to error," Life Sci. **22**, 1305–1360 (1978).

[21] W. H. Jefferys, "On the method of least squares," Astrophys. J. **85**, 177–181 (1980); **95**, 1299 (1988); "On the method of least squares. II," *ibid* **86**, 149–155 (1981); **95**, 1300 (1988).

[22] W. H. Jefferys, M. J. Fitzpatrick, and B. E. McArthur, "GaussFit—a system for least squares and robust estimation," Celestial Mech. **41**, 39–49 (1988). See also W. H. Jefferys, "Robust estimation when more than one variable per equation of condition has error," Biometrika **77**, 597–607 (1990).

[23] S. D. Christian and E. E. Tucker, "Least squares analysis with the microcomputer," Am. Lab. **14**, 642–646 (1984).

[24] D. York, "Least-squares fitting of a straight line," Can. J. Phys. **44**, 1079–1086 (1966).

[25] J. H. Williamson, "Least-squares fitting of a straight line," Can. J. Phys. **46**, 1845–1847 (1968). See also D. P. Chong, "Comment on 'Linear least-squares fits with errors in both coordinates,' by B. C. Reed [Am. J. Phys. **57**, 642–646 (1989)]," Am. J. Phys. **59**, 472–474 (1991), and B. C. Reed, "Linear least-squares fits with errors in both coordinates II. Comments on parameter variances," to be published in Am. J. Phys.

[26] E. E. Tucker, private communication (1990).

[27] D. R. Powell and J. R. Macdonald, "A rapidly convergent iterative method for the solution of the generalized nonlinear least squares problem," Comput. J. **15**, 148–155 (1972); **16**, 51 (1973).

[28] H. I. Britt and R. H. Luecke, "The estimation of parameters in nonlinear, implicit models," Technometrics **15**, 233–247 (1973).

[29] M. Clutton-Block, "Likelihood distributions for estimating functions when both variables are subject to error," Technometrics **9**, 261–269 (1967).

[30] J. G. Hust and R. D. McCarty, "Curve-fitting techniques and applications to thermodynamics," Cryogenics **7**, 200–206 (1967).

[31] J. P. Chandler, "On an iterative procedure for estimating functions when both variables are subject to error," Technometrics **14**, 71–76 (1972).
[32] A. Madansky, "The fitting of straight lines when both variables are subject to error," J. Am. Stat. Assoc. **54**, 173–203 (1959).
[33] R. H. Luecke, H. I. Britt, and K. R. Hall, "Comparison of two algorithms for least squares parameter estimation," Cryogenics **14**, 284 (1974).
[34] J. R. Macdonald, "Impedance spectroscopy: old problems and new developments," Electrochim. Acta **35**, 1483–1492 (1990).
[35] J. R. Macdonald and W. J. Thompson, "Strongly heteroscedastic nonlinear regression," to be published in Commun. Statistics Simulat. Comput. B **20**(4) (1991).

Deterministic chaos in the elastic pendulum: A simple laboratory for nonlinear dynamics

R. Cuerno, A. F. Rañada, and J. J. Ruiz-Lorenzo
Instituto de Física Fundamental y Departamento de Física Teórica I., Facultad de Ciencias Físicas, Universidad Complutense, 28040 Madrid, Spain

(Received 25 July 1990; accepted 7 May 1991)

The chaotic motion of the elastic pendulum is studied by means of four indicators, the Poincaré section, the maximum Lyapunov exponent, the correlation function, and the power spectrum. It is shown that for very low and very large energies the motion is regular while it is very irregular for intermediate energies. Analytical considerations and graphical representations concerning the applicability of KAM theorem are also presented. This system and the type of description used are very suitable to introduce undergraduate students to nonlinear dynamics.

I. INTRODUCTION

The elastic pendulum, although rather a simple mechanical system, combines a complex dynamical behavior with a wide applicability as a mathematical model in different fields of physics, such as nonlinear optics or plasma physics.[1] The techniques used to tackle this simple but at the time complex system range from perturbative studies (where parametric resonance has been found, due to the existence of energy transfer among the different modes)[1–3] to experimental studies that use stroboscopic techniques.[4] In any case it is clear that such an apparently uninteresting system with just 2 degrees of freedom displays a rich and varied dynamics. A method that is complementary to the above mentioned, and which has been widely used in the literature for other dynamical systems, is the numerical computation of several indicators able to characterize the kind of evolution one has for each set of parameters and initial conditions. However, each indicator alone can be misleading. We shall see later that the joint use of several indicators may greatly clarify the analysis of the evolution. Our system displays one kind of motion for a range of parameters, and a drastically different motion for other values. This feature enables us to compare the efficiency of standard methods in the numerical characterization of chaos.

We consider a Hamiltonian system, with N degrees of freedom, to be chaotic when the maximum number of dynamical variables in involution (i.e., with Poisson brackets equal to zero) is less than the number of degrees of freedom N. This is because an important theorem due to Liouville[5,6] states that when there are N conserved quantities in involution the solution of the equations of motion can be obtained by quadratures and the behavior is regular. Moreover, it is observed that when this is not the case the system behaves stochastically, at least some of the solutions being unstable. As it is not easy to prove that there are no such quantities, one has to resort to some indicators, four of which are studied in this paper. We will admit that there is chaos if several of these indicators show this to be the case.

In the next section we shall briefly describe our system. We will stress the lack of a sufficient number of conserved quantities for the system to be exactly solved. Then we consider some formal arguments at high and low energy regimes, which may explain numerically observed behaviors. They will tell us about the exact integrability of the system at those particular regimes. The applicability of the KAM theorem will also be examined and consistency with numerical results will be checked. In the third section the main results of the paper are presented. We characterize the motion by use of four different numerical indicators (Poincaré section, maximum Lyapunov exponent, correlation function, and power spectrum). This is done for examples of both regular and irregular types of motion. Finally, some conclusions are presented.

II. EQUATIONS OF MOTION

The Lagrangian of the plane elastic pendulum, in the Cartesian coordinates of Fig. 1, is

$$L = m(\dot{x}^2 + \dot{y}^2)/2 - mgy - (k/2)[(x^2+y^2)^{1/2} - l_0]^2, \tag{1}$$

where no approximation has been made; l_0 is the natural length of the pendulum, k is the spring elastic constant, m is the mass of the bob, and g is the gravitational acceleration on the Earth surface. The Euler–Lagrange equations of motion are

$$\ddot{x} = -\omega_s^2 x + g\lambda x/(x^2+y^2)^{1/2}, \tag{2}$$

$$\ddot{y} = -g - y\omega_s^2 + g\lambda y/(x^2+y^2)^{1/2}, \tag{3}$$

