Linear least-squares fits with errors in both coordinates

B. Cameron Reed
Department of Physics, Saint Mary's University, Halifax, Nova Scotia B3H 3C3, Canada

(Received 28 July 1987; accepted for publication 2 August 1988)

York's solution to the problem of linear least-squares fits with errors in both coordinates [D. York, Can. J. Phys. **44**, 1079 (1966)] is shown to be exact and not subject to the erroneous results that attempts to modify standard least-squares algorithms can produce. Detailed examples of the use of York's method are given; a FORTRAN implementation suitable for use on personal computers is available to interested parties.

I. INTRODUCTION

It is virtually certain that any physical scientist will at some time be faced with the task of determining the "best" straight line through data in the Cartesian plane. Students of physics are usually exposed to this "linear least-squares" problem as undergraduates in advanced laboratory courses, and the solution when the abscissa values are taken to be error-free is well known[1] and now involves no more than a few key strokes on a calculator. However, on the problem of least-squares fits involving errors in *both* coordinates, the textbooks largely fall silent. The intuitively obvious approach of defining the best-fit line as that which minimizes the sum of the squares of the *perpendicular* distances of the points from the line is suitable only when all points have equal weights in both coordinates. A fully general solution capable of admitting varying weights is necessary.

Deming[2] gave a general solution based on minimizing the sum

$$S=\sum\left[W(x_i)(x_i-X_i)^2+W(y_i)(y_i-Y_i)^2\right],\qquad(1)$$

where (x_i, y_i) are the observed points, (X_i, Y_i) are predicted values, and $W(x_i)$ and $W(y_i)$ are the weights in x and y for point $i(i=1,N)$, usually assumed to be the reciprocals of the squares of the measurement uncertainties. Deming simplified the problem by expanding in a Taylor series about assumed values of the slope, intercept, and adjusted points, dropping squared and higher terms. Unfortunately, neglect of the higher-order terms can lead to significant errors in some circumstances.

An exact, completely general solution to the linear least-squares problem, apparently first given by York,[3] seems not to be well known despite the frequency with which this problem arises in research and teaching circles. Unfortunately, York's article contains misleading statements that can lead to trouble in some circumstances. Apparently unaware of York's work, Orear[4] devised a solution based on the "effective variance method," which he proved to be exact for linear fits and applicable in an approximate fashion to nonlinear functions. Lybanon[5] (see also Orear[6]) pointed out that an iterative utilization of Orear's method would not lead to exact maximum likelihood parameter estimates, and that in some not uncommon cases no change in the parameter estimates will result from subsequent iterations! This problem can be safely circumvented, however, provided one utilizes Orear's method in the way discussed by Lybanon.

Given that readers may not be aware of the above problems, it is possible that many are utilizing erroneous algorithms. The purpose of this article is to resolve the misleading statements in York's article, to show that his method is an exact solution immune from the trap that modified least-squares algorithms can fall prey to, and to make available a simple FORTRAN implementation of York's method suitable for use on a personal computer.

The outline of this article is as follows: In Sec. II, I reiterate York's solution, draw attention to his misleading statements, and indicate an easier way to solve for the slope of the best-fit line. In Sec. III, I apply York's method to examples of well-correlated and poorly correlated data. Finally, in Sec. IV, I compare York's method to the Orear–Lybanon approach, show that they are equivalent, and argue that York's solution cannot fall prey to Lybanon's "trap."

II. YORK'S SOLUTION

Assuming that the function to be fitted is of the form

$$y=mx+c,\qquad(2)$$

York showed that the slope m is given by solving the equation

$$f(m)=m^3-3\alpha m^2+3\beta m-\gamma=0,\qquad(3)$$

where

$$\alpha=\left(2\sum\frac{W_i^2U_iV_i}{W(x_i)}\right)(3\delta)^{-1},\qquad(4)$$

$$\beta=\left(\sum\frac{W_i^2V_i^2}{W(x_i)}-\sum W_iU_i^2\right)(3\delta)^{-1},\qquad(5)$$

$$\gamma=\left(-\sum W_iU_iV_i\right)\delta^{-1},\qquad(6)$$

and

$$\delta=\sum\frac{W_i^2U_i^2}{W(x_i)},\qquad(7)$$

where $W(x_i)$ and $W(y_i)$ are the x and y weights to be assigned to the data points

The "overall" weights W_i are given by

$$W_i=W(x_i)W(y_i)/[m^2W(y_i)+W(x_i)]\qquad(8)$$

with

$$U_i=x_i-\langle x\rangle,\qquad(9)$$

$$V_i=y_i-\langle y\rangle,\qquad(10)$$

where

$$\langle x\rangle=\sum W_ix_i\left(\sum W_i\right)^{-1}\qquad(11)$$

and

$$\langle y\rangle=\sum W_iy_i\left(\sum W_i\right)^{-1}.\qquad(12)$$

The intercept c is given by

$$c = \langle y \rangle - m \langle x \rangle, \tag{13}$$

and the errors in the slope and intercept are given by

$$\sigma_m^2 = \frac{1}{N-2} \sum W_i (mU_i - V_i)^2 \Big(\sum W_i U_i^2 \Big)^{-1} \tag{14}$$

and

$$\sigma_c^2 = \Big[\sum W_i x_i^2 \Big(\sum W_i \Big)^{-1} \Big] \sigma_m^2, \tag{15}$$

respectively.

Finally, in terms of the W_i, the sum of squared residuals can be written as

$$S = \sum_{i=1}^{N} W_i (y_i - c - mx_i)^2. \tag{16}$$

York termed Eq. (3) the "least-squares cubic." This equation will have three roots, one of which is presumably the "correct" least-squares slope. The catch is that Eq. (3) is not really a cubic because of the implicit dependence of α, β, and γ on the slope m via the weights W_i given by Eq. (8). York advocated an iterative approach: An initial guess for m (say, via a no-errors solution) is provided for use in Eq. (8), Eq. (3) is solved and the "correct" root is used to update the weights in Eq. (8), and so forth until convergence is reached. Intuitively, one would expect this procedure to converge quickly, particularly for well-correlated data. We shall see that this is not the case, and that convergence does not necessarily imply that one has found the "correct" root.

Analytically, the three roots of Eq. (3) are given by

$$m_1 = \alpha + (C + D), \tag{17}$$

$$m_2 = \alpha - \tfrac{1}{2}(C + D) + \tfrac{1}{2}(C - D)\sqrt{3}i, \tag{18}$$

$$m_3 = \alpha - \tfrac{1}{2}(C + D) - \tfrac{1}{2}(C - D)\sqrt{3}i, \tag{19}$$

where

$$C = (B + \sqrt{A^3 + B^2})^{1/3}, \tag{20}$$

$$D = (B - \sqrt{A^3 + B^2})^{1/3}, \tag{21}$$

with

$$A = (\beta - \alpha^2) \tag{22}$$

and

$$B = -\tfrac{3}{2}\alpha\beta + \tfrac{1}{2}\gamma + \alpha^3. \tag{23}$$

York states (without proof) that Eq. (3) will have three real roots. This is not true in general, and will lead to disaster if one attempts to computerize his iterative method using only real variables. A further complication is that York states that it is usually the third root of the cubic [Eq. (19)] that is the correct one. While the experience of this author tends to confirm that wisdom, I find as well that in some instances the first few iterations may yield a *complex* third root from which one must carve off the real part as the seed value for subsequent iterations. These problems are illustrated in Sec. III.

With the advent of interactive personal computers, however, a much more straightforward solution is possible: Simply scan $f(m)$ for its real roots (zeros), bearing in mind that α, β, and γ are functions of m. With even a crude idea of the slope to start with, it is possible to pin down the roots very quickly to a number of significant figures greater than could possibly be justified by the input data. This intuitively appealing procedure avoids the difficulties of manipulating (and interpreting) complex quantities.

In Sec. III, I illustrate application of the above "iteration" and "root-finding" approaches to York's method to two disparate cases, one of well-correlated real data and one of poorly correlated fictitious data.

III. EXAMPLES

If one elects to pursue the iterative path in applying York's method, it is essential to realize at the outset that due to the complexity of the coefficients in Eq. (3) it is impossible to give any general guidelines as to when one might expect imaginary roots, or as to when the third root will or will not be the "correct" one. Armed on the other hand with the root-finding method, however, it is merely necessary to evaluate $f(m)$ over suitable ranges in m until one has isolated the zeros. It is instructive to compare these two methods.

To begin with, I developed two separate FORTRAN codes for use on my IBM-PC. Both commence by giving the operator a trial slope based on a conventional, no errors, least-squares fit. The "iterative" program, utilizing complex arithmetic, then computes and prints the roots of the "least-squares cubic." The operator selects the real part of one of the roots as the input slope for the next pass. Interation proceeds as long as the operator desires. In the "root-finding" program, one inputs a range (and increment) of slopes for which $f(m)$ is to be evaluated. Once the general location of a root has been established, subsequent passes over smaller ranges with smaller increments can be used to pin down the root with arbitrarily great accuracy. In practice, this procedure is very rapid.

A. Example I: Well-correlated data

Table I gives (x,y) values for 27 points, which are plotted in Fig. 1. All points were assumed to have uncertainties of 0.01 in both x and y, and weights were assigned as the squares of the reciprocals of the uncertainties. These data derive from a real, physical situation: calibrating the colors of globular star clusters as a function of their spectral types.[7] The abscissa represents the difference between the ultraviolet and yellow light magnitudes of the clusters and the ordinate represents the difference between the yellow and infrared magnitudes. The slope given by a conventional (minimizing y residuals) least-squares fit is 0.931. With this value used as a first guess for m, the iterative solution yields two conjugate complex roots (the first and third), $1.383 \pm 0.098i$, and one real root (the second), -0.904—

Table I. Data for Example I.

x	y	x	y	x	y
0.89	0.67	1.01	0.77	0.88	0.79
1.00	0.64	0.86	0.73	0.92	0.77
0.92	0.76	0.85	0.64	0.92	0.70
0.87	0.61	0.88	0.62	1.01	0.88
0.90	0.74	0.84	0.63	0.88	0.62
0.86	0.61	0.79	0.57	0.92	0.80
1.08	0.77	0.88	0.66	0.96	0.74
0.86	0.61	0.70	0.53	0.85	0.64
1.25	0.99	0.81	0.46	1.04	0.93

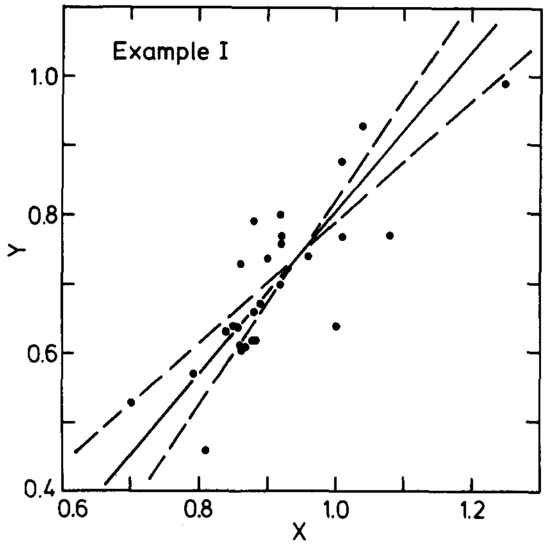

Fig. 1. Linear least-squares fit for the data in Table I. The solid line corresponds to best-fit parameters $m = 1.167 \pm 0.308$ and $c = -0.365 \pm 0.291$. The dashed lines show the fit for the combinations of greatest and least slopes and intercepts. All points were assumed to have uncertainties of ± 0.01 in both x and y.

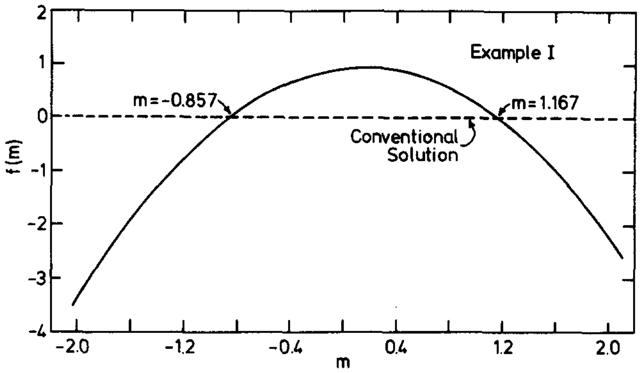

Fig. 2. Behavior of $f(m)$ for the data in Table I.

nowhere near what one might expect from looking at the data. As we are interested in a real root, intuition suggests the second root as the one to select for subsequent iterations. This leads to convergence following 17 iterations to two complex roots, $1.360 \pm 0.190i$, and one real root (again, the second), -0.857. While convergence to a real root has been achieved, it is obvious that this root cannot possibly be the correct best-fit slope.

The only course of action here is to investigate the consequences of selecting one of the other roots as the seed value for the second and subsequent iterations. The choice is irrelevant as both the first and third roots have the same real component. This leads to *three real, unequal roots* following the second iteration. Continuing with the first root as the input slope for subsequent iterations leads always to three real roots; however, there is no indication of convergence following 30 iterations. On the other hand, if the *third* root is selected following the *second* iteration, convergence to three real roots, 1.764, -1.068, and 1.167, is achieved after nine iterations. The *third root* (consistent with York's statement) is apparently correct, and has an rms deviation of 0.072. This result is shown as a solid line in Fig. 1.

Clearly, the appearance of a real (complex) root following the first iteration does not imply that root to be the correct (incorrect) one. *Complex roots can arise from well-correlated data; one must be prepared in the iterative approach to look at the results of selecting various roots for subsequent iterations.*

The root-finding approach quickly yielded the same two real roots as the iterative approach: -0.857 ± 0.255 and 1.167 ± 0.308, with sums of squared residuals of 5968.4 and 578.0, respectively. The behavior of $f(m)$ is illustrated in Fig. 2.

While it is immensely satisfying that both implementations of York's method yielded the same real roots, the interpretation of the negative root may not at first sight be clear. The answer is that the two roots correspond, respectively, to *maximizing* and *minimizing* Eq. (1). York established the "least-squares cubic" by differentiating Eq. (1), setting the result to zero, and solving by the method of undetermined multipliers. Solving for the roots of the "least-squares cubic" yields only those values of m for which S is an extremum; to establish which root is the correct one requires calculation of S once m and c are known. This leaves open the question of possible other *local* extrema for S; of course, in any "real" situation, one is likely to have at least a crude idea of what the "true" slope is. Also, it should be noted that the roots corresponding to global extrema in S should yield perpendicular fits within their errors; in the case of equal uncertainties in both x and y, the product of these two roots should be -1. This is the case in the above example.

A parenthetic remark concerning this example is appropriate at this point. With the fit shown in Fig. 1, 3 of the 27 points are deviant by more than 8 s.d. As one would expect if $S \sim 25$ for 27 points, then either the assumption of a linear relation between x and y is invalid, or there are systematic errors present that mask the statistical errors. Regardless of these problems, this example serves to illustrate some of the situations one can encounter when dealing with errors in both coordinates.

Finally, this example should not be construed to imply that iterative convergence will be achieved even if the real part of an initially complex root is selected as the seed value for subsequent iterations. This is demonstrated in the next example.

Table II. Data for Example II.

x	y
1.333 ± 2.469	− 1.367 ± 0.297
− 1.009 ± 6.363	7.232 ± 4.672
9.720 ± 6.045	− 0.593 ± 2.014
− 2.079 ± 4.061	7.124 ± 0.022
8.920 ± 5.325	0.468 ± 6.868
− 0.938 ± 5.865	8.664 ± 2.834
10.94 ± 3.993	5.854 ± 4.647
5.138 ± 3.787	13.35 ± 4.728
11.37 ± 3.693	4.279 ± 2.274
9.421 ± 4.687	11.63 ± 4.659

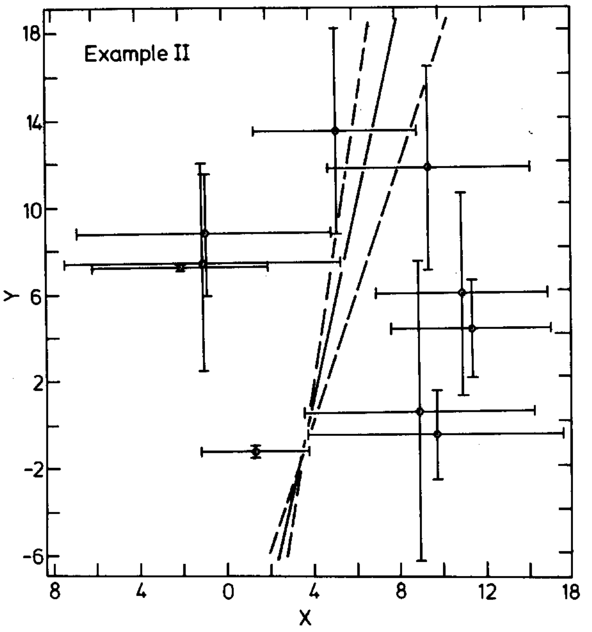

Fig. 3. Least-squares fit for the data in Table II. Points were weighted inversely as the squares of the uncertainties. Here, $m = 4.544 \pm 1.576$ and $c = -17.483 \pm 5.324$.

B. Example II: Poorly correlated data

Table II and Fig. 3 show ten poorly correlated data points and their associated uncertainties, created with a random number generator. A conventional no-errors solution gives $m = -0.157$. The first iteration of the least-squares cubic yields the third root to be real (0.010) and the first two to be complex ($-1.636 \pm 1.028i$). Consistently selecting of the real part of the third root as the input slope for subsequent iterations leads to an endlessly repeating pattern: Iterations n, $n+1$, and $n+2$ produce roots identical to iterations $n+3$, $n+4$, and $n+5$, respectively. All three roots of iteration n are real (2.102, -2.006, 0.170) while only the third root of iteration $n+1$ is real (0.011) and only the first root of iteration $n+2$ is real (0.001 71). Further experimentation showed that if the second root is selected neither a repeating pattern nor any clear trend to convergence is evident after 30 iterations. Convergence is finally achieved on selecting the real part of the first root; after 12 iterations, it converges to 0.001 66 ($S = 833.4$) and the other two roots to $-2.192 \pm 0.216i$. Curiously, it was found that if the third (real) root was selected following the first iteration and the first root thereafter, convergence to the same result is achieved in only four iterations!

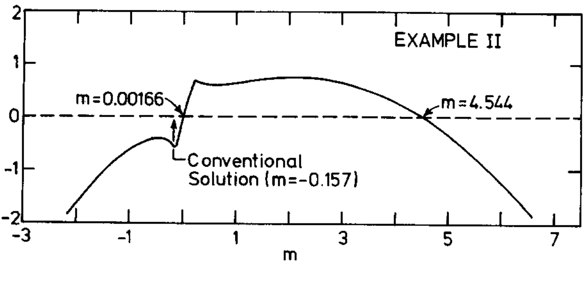

Fig. 4. Behavior of $f(m)$ for the data in Table II.

With the root-finding algorithm, one can quickly zero in on the root at $m = 0.001\,66$. If this is indeed the correct best-fit slope, we should expect to find a corresponding maximum S slope at $-1/0.001\,66^{\sim} -602$. No such root is evident. After some experimentation, however, a second root does turn up at $m = 4.544 \pm 1.576$ with $S = 13.96$. This root—missed completely by the iterative approach—is the true best-fit slope. This result is shown as a solid line in Fig. 3; the behavior of $f(m)$ is illustrated in Fig. 4.

IV. COMPARISON WITH OREAR'S METHOD

Given that Lybanon has pointed out potential "traps" in Orear's general solution to the linear least-squares problem, it is important to compare the York and Orear solutions. Is York's exact? Are they equivalent? If so, is York's open to the same criticism, and which is to be preferred computationally? The answers to these questions prove to be yes, yes, no, and York, respectively.

That York and Orear address the same mathematical question is evident by inspecting their equations for the sum of the weighted squared residuals [Eq. (7) in both cases—Eq. (1) or (16) in the present article]. They are identical provided York assigns weights as the inverse square of the measurement uncertainties; Orear's effective weights $1/\delta j^2$ [his Eq. (5)] are identical to York's weights W_i [Eq. (8) above] in this case. The crux of Lybanon's critique is Orear's assertion that a standard least-squares algorithm can still be used provided one utilizes his effective weights. Such algorithms usually rely on a matrix inversion to generate the fit parameters. Presumably, one would start off with an initial guess for the slope, compute the effective weights, run the standard least-squares program to get fit parameters that are used to update the weights, and so on. The problem is that this procedure will not yield exact least-squares parameters because the standard least-squares programs ignore the dependence of the weights on the unknown fit parameters. Lybanon gives an appealing simple example: If the uncertainties in all of the x and y values are equal (but not necessarily the same for each dimension), then over one iteration the weights can be factored out of the expression for S and a least-squares algorithm is left with a regression of y on x with equally weighted points, the result being that no change will result from "updating" the weights. This is precisely the case of Example I in Sec. III, yet the "York-iterative" method clearly gave evolving slopes and agreed with the root-finding method. What makes it immune from Lybanon's critique? The answer is that nowhere does it utilize a standard least-squares algorithm. In the case of equal weights, it is easy to show that the weights *cannot* be factored out of β and γ; either iteration or root finding remains sensitive to changes in the weights caused by varying the slope in all circumstances. Curiously, in the case of equal weights, the least-squares cubic boils down to a *true* cubic equation that could be solved directly.

In summary, the York and Orear formulations yield identical exact results *provided* one weights in inverse-square proportion to the uncertainties and *minimizes* Orear's expression for the effective variance S. York's formulation can be solved iteratively, by root finding, or by minimizing his expression for S [Eq. (16) above]. However, there is a significant advantage to root finding: Minimizing S is a two-dimensional problem (both slope and intercept appear), whereas only m appears in Eq. (3).

York's method also has the advantage of admitting different weighing schemes should one so desire.

V. CONCLUDING REMARKS

In this article, I have pointed out some misleading statements given by York in his general solution to the linear least-squares problem. York's solution is exact, equivalent to that published by Orear, and free of the potential difficulties pointed out by Lybanon. Finding the roots of York's "least-squares cubic" is simpler than minimizing the effective variance. Finally, it is worthwhile reiterating that standard least-squares algorithms cannot be modified to yield correct (exact) results for the problem of linear least-squares fits with uncertainties in both coordinates. The irony in this is that the correct solution is almost as simple.

The author will be pleased to make available a copy of his root-finding algorithm to any interested parties.

ACKNOWLEDGMENTS

I wish to thank Dave Turner for initially bringing York's article to my attention, and Steve Shawl and Jim Hesser for encouraging me to explore its subtleties. I am grateful to an anonymous referee for comments that led to significant improvements in both my own understanding of this problem and the presentation of this article. Dr. David Swingler provided a number of valuable suggestions and comments. I am grateful to Randall Brooks for preparing the figures, and to Karen Rieber and Susan Storring for word processing. This work was supported by National Sciences and Engineering Research Council of Canada Operating Grant A8580.

[1] P. R. Bevington *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1969), Chap. 6.

[2] W. E. Deming, *Statistical Adjustment of Data* (Wiley, New York, 1943).

[3] D. York, Can. J. Phys. **44**, 1079 (1966).

[4] J. Orear, Am. J. Phys. **50**, 912 (1982).

[5] M. Lybanon, Am. J. Phys. **52**, 276 (1984).

[6] J. Orear, Am. J. Phys. **52**, 278 (1984).

[7] B. C. Reed, J. E. Hesser, and S. J. Shawl, Pub. Astron. Soc. Pacific **100**, 545 (1988).

Some simple experimental studies using a passive cavity coupled to a He–Ne laser cavity for practice in a quantum electronics laboratory

Adrian Gh. Podoléanu and Ion M. Popescu
Department of Physics, Bucharest Polytechnical Institute, Spl. Independentei 313, Bucharest, Romania

(Received 23 November 1987; accepted for publication 29 August 1988)

Some quantum optics experiments about self-locking and mode-locking operation, optical bistable operation, and low-absorption monitoring are described using a versatile, inexpensive experimental setup. These experiments are easy to implement with a He–Ne laser coupled to a passive cavity to which a few simple devices have been added.

I. INTRODUCTION

Progress in quantum electronics has been quite rapid in the last 20 years. Therefore, besides classical optical experiments, laboratory courses should introduce experiments connected to the specific aspects of laser interaction with matter, such as multiphoton absorption, harmonic generation, mode locking, optical bistability, etc.

Other articles in this Journal have already mentioned some relatively simple laboratory experiments on optical cavities and laser mode beats,[1–4] self-locking,[1] construction of a dye laser[5] as well as on optogalvanic spectroscopy.[6]

The present article describes a versatile laboratory setup that may be used for a great variety of quantum electronics experiments. The main part of this setup consists of a He–Ne laser coupled to a passive cavity. By adding a few inexpensive devices to this setup, different configurations may be obtained to illustrate self-locking, mode locking, optical bistable operation, and low-absorption coefficient measurement possibilities.

II. MULTIMODE LASER BEHAVIOR

Lasers normally oscillate simultaneously in a number of resonator modes of different frequencies. Neglecting the transverse mode effects, the mode frequency is determined by the number of half-wavelengths of the radiation contained in a cavity of length L.

If the modes are considered equispaced, the frequency difference between adjacent modes is given by the relation

$$\Delta f_{\mathrm{ax}} = (1/2L)(c/n), \tag{1}$$

where n is the effective refractive index within the cavity and c is the light speed *in vacuo*. The relation (1) is approximate, being complicated by the fact that c/n is affected by dispersion. Since the velocity c/n depends on frequency, the modes will not actually be equally spaced in frequency.

Every mode satisfies the Maxwell wave equation. In the free mode of operation, these equations are independent. If a certain nonlinearity occurs, this will trigger the coupling