

The Analysis of Ligand-Binding Data with Experimental Uncertainties in the Independent Variables¹

MICHAEL L. JOHNSON

Biophysics Program, Diabetes Research and Training Center, and Department of Pharmacology, University of Virginia School of Medicine, Charlottesville, Virginia 22908

Received December 13, 1984

The method of experimental data analysis known as least-squares requires several inherent assumptions to be met in order for the analysis to be statistically correct. In particular, it must be assumed that the experimental uncertainties exist only on the dependent variables. Least-squares is often used for applications where this assumption is not satisfied. An alternative method of data analysis circumvents the assumption that no experimental uncertainty exists in the independent variables. This method, known as maximum likelihood, will produce statistically correct results in cases where experimental uncertainty exists on both the dependent and independent variables. The method can easily be generalized to include cases where the dependent and independent variables are cross-correlated. The method can also be generalized to include non-Gaussian distributions of experimental uncertainties. The examples presented are simulated applications to ligand-binding problems. The general method is, however, applicable to a wide range of problems in biochemistry. © 1985 Academic Press, Inc.

KEY WORDS: numerical methods; least-squares; ligand binding.

The statistically correct application of any curve-fitting procedure used to obtain estimates of derived parameters requires that a number of assumptions be satisfied. For instance, the parameter estimation procedure must be tailored to suit the shape and magnitude of the particular distribution of random experimental uncertainties which are inherent in both the dependent variable (the ordinates) and the independent variables (the abscissas). The parameter estimation procedure must also consider any correlations between the experimental uncertainties of the various dependent and independent variables.

The parameter estimation procedure commonly known as nonlinear least-squares makes a number of limiting assumptions about the distributions of experimental uncertainties. In particular, the least-squares method assumes (i) that negligible experi-

mental uncertainty exists in the independent variables, (ii) that the experimental uncertainties of the dependent variables are Gaussian in their distribution with a mean of zero; and (iii) that no correlation exists between the experimental uncertainties of the dependent and independent variables. Any application using least-squares where these assumptions are not met will yield incorrect answers. The magnitude of the errors so introduced is impossible to predict a priori, because it is a function of the particular data points and their experimental uncertainties as well as the functional form of the equation being fitted. The assumptions cannot be overcome by an appropriate weighting of the data.

The method of least-squares is often used for applications where these assumptions are not met. For example, the common method for determining a "standard curve" will usually neglect the experimental uncertainties of the independent variables (i.e., the x axis). In column chromatographic studies it is

¹ This work was supported in part by National Institutes of Health Grants GM28929, AM30302, and AM22125.

common to generate a standard curve of elution volume vs log molecular weight for a number of standard proteins. The molecular weight estimates of the standard proteins are not known without experimental uncertainties. In electrophoresis experiments log molecular weight is plotted against electrophoretic mobility. In this case, electrophoretic mobility should not be considered without its relative experimental uncertainty.

Other "plots" which are commonly used in biochemical literature which violate one or more of the assumptions of least-squares include Hill plots, Scatchard plots, and double-reciprocal plots.

The primary purpose of this work is to describe a method of parameter estimation which allows for experimental uncertainties in the independent variables. The method, as presented, still assumes that the experimental uncertainties follow a Gaussian distribution and are independent of each other. However, the generalization of this method to include cross-correlated non-Gaussian distributions is also discussed.

NUMERICAL METHODS

A parameter estimation procedure takes an equation of an assumed functional form and a set of data and generates a new function called a NORM, which shows a maximum or minimum when the parameter values, the desired "answers," show the highest probability of being correct. For the standard least-squares technique, this NORM of the data is given by

$$\text{NORM}(\alpha) = \sum_{i=1}^n \left[\frac{Y_i - G(\alpha, X_i)}{\sigma_i} \right]^2 \quad [1]$$

where α is any vector of parameters for an arbitrary function G and n data points (X_i, Y_i) , with each Y_i having a unique experimental uncertainty (standard error) of σ_i . The maximum likelihood estimate of the parameters, α , will correspond to a minimum of the NORM in this case.

A number of assumptions are implicit in the derivation of this least-squares NORM and must also be considered essential to the method derived in this paper (1). It must be assumed that there are enough data points to give a random sampling of the experimental uncertainty, and that the function, G , correctly describes the phenomenon occurring. If Gaussian-distributed random experimental uncertainty is assumed on both the ordinate and the abscissa, and that these experimental uncertainties are independent of each other, the statistically correct NORM will be similar to Eq. [1] (see Eq. [5] below). With the other previously mentioned assumptions it can be shown that the probability, P_i , for observing a particular data point (X_i, Y_i) at any value of the parameters, α , is proportional to

$$P_i(\alpha) \approx \frac{1}{2\pi\sigma_{X_i}\sigma_{Y_i}} \exp\left[-\frac{1}{2}\left[\frac{Y_i - G(\alpha, \bar{X}_i)}{\sigma_{Y_i}}\right]^2\right] \\ \times \exp\left[-\frac{1}{2}\left[\frac{X_i - \bar{X}_i}{\sigma_{X_i}}\right]^2\right] \quad [2]$$

where σ_{X_i} and σ_{Y_i} represent the standard deviations of the Gaussian distributed random experimental uncertainty at the particular data point and \bar{X}_i is the "optimal" value of the independent variable. In order to derive Eq. [2], and as a consequence of Eqs. [3]–[5], it has been assumed that the σ_{X_i} and σ_{Y_i} are independent of each other. It has not been assumed that they have any relationship, such as constant coefficients of variation, between σ_{X_i} and X_i or σ_{Y_i} and Y_i . The probability of making a series of measurements at n independent data points is then proportional to

$$P(\alpha) \approx \prod P_i(\alpha) \approx \left[\prod \left[\frac{1}{2\pi\sigma_{X_i}\sigma_{Y_i}} \right] \right] \\ \times \exp\left[-\frac{1}{2} \sum \left[\frac{Y_i - G(\alpha, \bar{X}_i)}{\sigma_{Y_i}} \right]^2\right] \\ \times \exp\left[-\frac{1}{2} \sum \left[\frac{X_i - \bar{X}_i}{\sigma_{X_i}} \right]^2\right] \quad [3]$$

where the product and summation are taken for each of the n data points with subscript i . This equation can be reorganized to the form:

$$P(\alpha) \approx \left[\prod \frac{1}{2\pi\sigma_{X_i}\sigma_{Y_i}} \right] \times \exp \left[-\frac{1}{2} \sum \left[\left[\frac{Y_i - G(\alpha, \bar{X}_i)}{\sigma_{Y_i}} \right]^2 + \left[\frac{X_i - \bar{X}_i}{\sigma_{X_i}} \right]^2 \right] \right]. \quad [4]$$

The maximum likelihood estimates for the parameters, α , with the current assumptions will be those values of α which maximize the probability given by Eq. [4]. This can be accomplished by minimizing the summation in the exponential term in Eq. [4]. The NORM to minimize is then

$$\text{NORM} = \sum \left[\left[\frac{Y_i - G(\alpha, \bar{X}_i)}{\sigma_{Y_i}} \right]^2 + \left[\frac{X_i - \bar{X}_i}{\sigma_{X_i}} \right]^2 \right]. \quad [5]$$

This is an extension of the least-squares NORM in Eq. [1], to include the possibilities of experimental uncertainties in the independent variables.

This new statistical NORM, Eq. [5], can be minimized by a recursive application of a curve-fitting algorithm such as Nelder-Mead (2,3). Equation [5] can be written as

$$\text{NORM} = \sum_{i=1}^n D_i^2 \quad [6]$$

where D_i is the weighted distance between the given data point (X_i, Y_i) and the point of closest approach to the fitted line ($\bar{X}_i, G(\alpha, \bar{X}_i)$).

The minimization of this norm can be performed by a nested, or recursive, minimization procedure. The parameter estimation procedure to evaluate the values of α is the standard Nelder-Mead simplex algorithm (2,3). An initial estimate of α is arbitrarily

chosen. This estimate is employed to calculate the D_i^2 at each data point. The D_i^2 values are then used to predict new values for the parameters being estimated, α . This cyclic process is repeated until the values do not change within some specified limit. An inconvenience arises in the evaluation of D_i^2 at each data point and iteration because it requires the value of \bar{X}_i . This value \bar{X}_i is the value of the independent variable, X_i , at the point of closest approach of the function to the particular data point evaluated at the current estimate of the parameters, α . The evaluation of this weighted distance of closest approach, D_i , includes the relative precision of the data point, σ_{X_i} and σ_{Y_i} , as per Eq. [5]. This implies that the values of \bar{X}_i will be different for each iteration. Because of this, these values of \bar{X}_i must be reevaluated for each iteration.

If the original parameter estimation procedure is carefully developed, it can be used recursively to evaluate \bar{X}_i . That is, at each iteration and data point of the parameter estimation procedure the routine calls itself to evaluate D_i^2 . For this recursive application of the algorithm, the parameters being estimated are now the \bar{X}_i , and the function being minimized is D_i^2 . The values of X_i or the previous values of \bar{X}_i are used as starting values for this iterative process. Function minimization algorithms such as Marquardt-Levenberg, Gauss-Newton, and steepest descent cannot be used for this purpose since they make assumptions about the functional form of the statistical NORM (1).

It should be noted that this method is not restricted to a two-dimensional problem, Y vs X . As with the least-squares NORM described in Eq. [1], all that is required to expand this method to include multiple dependent and multiple independent variables is to consider each of the Y_i, X_i and \bar{X}_i 's as vectors instead of scalars, and add the appropriate additional terms in the summations.

Furthermore, an equation analogous to Eq. [4] can be generated to include non-Gaussian distribution of experimental uncer-

tainties and/or any cross-correlation between the various dependent and independent variables. Then this equation is used as the NORM to be maximized to obtain the maximum likelihood parameter estimates. In a similar manner an additional term can be included in Eq. [5] to allow for the possibility of cross-correlation between the dependent and independent variables.

It is worth noting that this maximum likelihood approach allows for an extremely liberal choice of dependent and/or independent variables and the form of the fitting equation. For example, experimental data from the measurement of hormone binding to cell surface receptors usually consist of a series of measurements of amount bound vs total added hormone concentration. The most convenient way to formulate the fitting function, G , is as amount bound vs the free, not the total, hormone concentration. Some authors assume that the total added hormone concentration is equal to the free concentration. Some other authors calculate the free hormone concentrations as the total minus the bound hormone concentration. Neither of these approaches is statistically sound (1,4). A few authors take the more complex approach of fitting bound vs total hormone concentration (1,4). This approach involves evaluating the free concentration as the numerical root of a conservation of mass equation relating total and free hormone concentrations and the current best estimates of the fitting parameters, i.e., binding constants and capacities. This procedure does not allow for experimental uncertainties in the total hormone concentration.

The maximum likelihood method allows a different approach. The function G can be written for two dependent variables, total and bound hormone concentrations, as a function of the free hormone concentration. The free hormone concentration can be assumed to be any reasonable value with its standard error being plus or minus infinity. The algorithm will then find the optimal value of \bar{X}_1 , in this case the free concentration,

which best describes the dependent and independent variables, here the total and bound hormone concentrations, and their experimental uncertainties. This optimal value is then used for the calculation of the standard NORM. The net effect of this approach is that the data can be fit as a function of a quantity which was never measured, yet the procedure is still statistically correct!

If the experimental uncertainties are independent and Gaussian then the joint confidence intervals of the derived parameters can be determined by searching for combinations of the parameters which yield values of the NORM, Eq. [5], which are increased by a multiplicative factor proportional to the desired F statistic (1,5). This F -statistic (variance ratio) value is uniquely determined by the desired confidence probability and the number of degrees of freedom. The Hessian matrix which is required for this procedure was evaluated by the method outlined in the appendix to the original Nelder and Mead paper (3).

The technique presented here is a generalization of a technique presented by Acton (6) for use with straight line data. Acton's method allows for covariance between the independent and dependent variables, but does not allow for multiple independent or dependent variables. Furthermore, the method presented by Acton can only be used for a fitting function which is a straight line. The generalization I am presenting allows for multiple independent and dependent variables, nonlinear fitting functions, and a method of evaluating joint confidence intervals of the determined parameters.

The simulated data which I used to test this method included Gaussian distributed pseudorandom noise which was generated by averaging 12 evenly distributed random numbers over a range of ± 0.5 . These numbers were obtained from the Digital Equipment Corporation RT-11 Fortran 77 library function RANDU.

Each of the tests of the maximum likelihood method includes a comparison of the

same simulated data analyzed by a standard weighted nonlinear least-squares technique. The least-square method which I utilized is a modification of the Gauss-Newton procedure and has been described elsewhere (1,5).

TEST EXAMPLES AND RESULTS

To demonstrate the functionality of this method, I present two examples of its use. The examples are simulations of ligand-binding problems with either one or two classes of independent binding sites. In both cases, the number of binding sites in each class and the binding affinity of each class is assumed to be unknown.

Gaussian-distributed pseudorandom noise was superimposed on each set of data. In order to reduce the possibility of inadvertently using a nonrandom set of noise, and thus biasing the results, each calculation was performed 10 times with different sets of random noise.

The first example simulates an experimental system with a single class of binding sites with a binding affinity, K_a , of 10^5 M^{-1} and a maximal bound, B_{\max} , of 10^{-3} M . Ten logarithmically spaced data points were simulated over a concentration range of 10^{-6} to 10^{-4} M . This concentration range and affinity correspond to a fractional saturation ranging between 10 and 90%. These data were then perturbed with multiple sets of pseudorandom noise.

The simulated error for the dependent variable, the amount bound, is 10% of the actual amount bound. The corresponding error in the independent variable, the free concentration, is 7% of the free concentration, i.e., a constant coefficient of variation. An example of one such set of data is presented in Fig. 1. The ellipses in Fig. 1 were generated such that their major and minor axes correspond to one standard deviation experimental uncertainty.

For this first example, the desired parameters, the vector α , are the binding affinity, K_a , and the maximal bound, B_{\max} . The func-

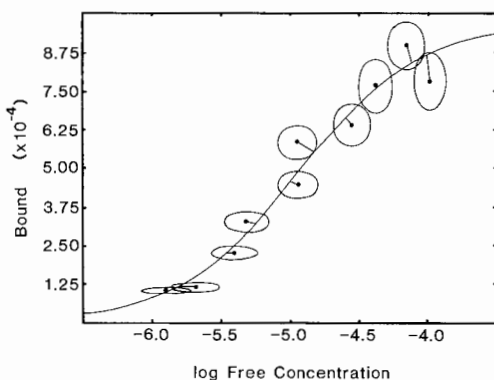


FIG. 1. Simulated data representing an experimental ligand-binding system with a single class of binding sites. The binding affinity, K_a , is 10^5 M^{-1} and the maximal bound, B_{\max} , is 10^{-3} M . The ellipses represent the Gaussian-distributed pseudorandom experimental uncertainty corresponding to 1 SD on both the ordinate and the abscissa. The curve is the calculated "best" fit of the data to the function by the maximum likelihood method of analysis. The short lines connecting the data points to the curve are the D_i 's in Eq. [6] whose sum of squares is being minimized by the analysis procedure.

tion G relates the amount bound, Y_i , with the free concentration, X_i , and α as

$$\text{Bound} = G(\alpha, X_i) = B_{\max} \frac{K_a X_i}{1 + K_a X_i} \quad [7]$$

The implementation of the maximum likelihood method presented here determines the logs of B_{\max} and K_a rather than the actual values. By allowing the logarithms of B_{\max} and K_a to assume any real value, I am generating a number system in which the actual values of B_{\max} and K_a are restricted to physically meaningful (positive) values.

The data were analyzed 10 times with separate sets of random noise by both a standard nonlinear least-squares technique and the maximum likelihood method which I am presenting. The average values and standard deviations of $\log K_a$ and $\log B_{\max}$ as determined by each procedure can be compared to measure the reliability and functionality of the analysis methods. The values determined by the maximum likelihood method were $\log K_a = 5.01 \pm 0.07$ and $\log B_{\max} = -3.01 \pm 0.03$. I then analyzed

exactly the same data with its superimposed pseudorandom noise by a weighted least-squares method. The values of σ_{Y_i} were used as a weighting factor for the analysis, i.e., the experimental uncertainties in the x axis were ignored. The values obtained were $\log K_a = 4.98 \pm 0.12$ and $\log B_{\max} = -3.01 \pm 0.06$. The average values as determined by both methods are excellent. However, the reproducibility of the values, measured by an F test, is decidedly better for the maximum likelihood method ($P \sim 95\%$ for each of $\log K_a$ and $\log B_{\max}$).

The calculated "best" fit of the data presented in Fig. 1 to the function by the maximum likelihood method of analysis is shown by the curve in Fig. 1. The short lines connecting each data point, which is the center of its ellipse, with the curve are the corresponding distances D_i in Eq. [6], whose sum of squares is being minimized by this maximum likelihood method. These distances, D_i , are perpendicular neither to the best curve nor to either axis. They are determined by a combination of the slope of the best curve and the relative errors in the dependent and independent variables. In Fig. 1, when the error is predominately in the x axis, these lines are nearly horizontal; when the error is predominately in the y axis, these lines are nearly vertical, and when the error is nearly equal, the line is almost perpendicular to the best curve.

The second example I chose to test is a simulation of a ligand-binding system with two classes of binding sites which differ in affinity by a factor of 5, and have equal binding capacities: i.e., K_a 's of 10^5 M^{-1} and $5 \times 10^5 \text{ M}^{-1}$ and B_{\max} 's of 10^{-3} M . Ten data points were simulated with a concentration range of 6×10^{-8} to $3 \times 10^{-4} \text{ M}$ and logarithmic spacing. The magnitude of the simulated experimental uncertainty was assumed to be the same as for the first example. Again, the data were analyzed 10 times with differing sets of superimposed pseudorandom noise. One of these analyses is shown in Fig. 2.

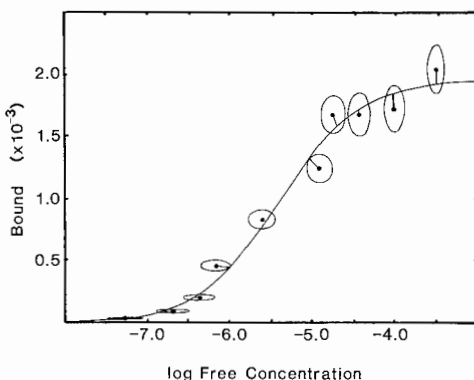


FIG. 2. These simulated data show a ligand-binding system with two independent classes of binding sites. K_a 's are 10^5 M^{-1} and $5 \times 10^5 \text{ M}^{-1}$ and B_{\max} is 10^{-3} M . Experimental uncertainty was imposed the same way as in Fig. 1, and the curve was calculated by the maximum likelihood analysis method.

The numerical values obtained by using maximum likelihood and weighted least-squares methods to analyze this second group of stimulated data are shown in Table 1. The maximum likelihood method gives more consistent results for both the high ($P > 99.9\%$)- and low ($P \sim 95\%$)-affinity classes of sites. The region of the data which contributes most to the determination of the low-affinity site is the upper half of the saturation curve. It is this upper portion of the curve which comes the closest to satisfying the least-squares criterion of negligible error in the x axis. Therefore, it is expected that the low-affinity class of sites would be evaluated with reasonable precision by the least-

TABLE I
ANALYSIS OF SIMULATED DATA FOR TWO CLASSES
OF BINDING SITES

	Correct values	Maximum likelihood values	Least-squares values
Log K_1	5.00	4.92 ± 0.23	4.90 ± 0.56
Log $B_{\max 1}$	-3.00	-2.97 ± 0.09	-2.80 ± 0.17
Log K_2	5.70	5.71 ± 0.08	6.32 ± 1.93
Log $B_{\max 2}$	-3.00	-3.01 ± 0.05	-4.15 ± 1.15

squares method. Conversely, the error on the lower portion of the curve is almost totally horizontal, which dramatically violates the error distribution assumption of the least-squares analysis method. Therefore, it is expected that, for this example, the high-affinity class of sites cannot be evaluated by least-squares.

The method of determining joint confidence intervals of the parameters was also tested for these two simulated examples. For each set of data, the maximum likelihood algorithm is capable of predicting a 1 SD confidence interval. In an ideal case, the average span of these predicted confidence intervals should be equal to twice the standard deviation of the determined parameter values averaged over the 10 different sets of pseudorandom noise. As measured by this criterion, the joint confidence intervals of the determined parameters are overestimated by approximately 75%. It has been previously shown that the expected joint confidence intervals of nonlinear parameters will be both asymmetrical and highly correlated (5). Consequently, the determination of a symmetrical standard deviation for these parameters from the results obtained with multiple sets of random noise is not optimal.

DISCUSSION

The first test case was specifically chosen as a simple two-parameter problem. The cross-correlation² between the two unknown parameters was relatively low: ≈ 0.87 . The second test was chosen because it poses a particularly difficult data analysis problem. The cross-correlation between the parameters

for the second test was very high: ≈ 0.99 . The second test represents a class of problems commonly referred to as mathematically ill-posed problems.

The convergence properties of the maximum likelihood method appear to be very good. In 2 of the 10 calculations performed by the least-squares technique on the second test example, the analysis would not converge, even when the correct answers were used as initial starting values. The maximum likelihood method converged on all of these in about 30 s on our microprocessor (DEC LSI-11/73) with no apparent problem.

The mathematical procedure presented here was specifically formulated to treat data analysis problems with experimental uncertainties in both the independent and dependent variables. Experimental uncertainties in the independent variables have previously been treated by a number of methods. The most prevalent method to treat these types of experimental errors is to ignore them. Some investigators minimize the sum of the squares of the perpendicular distance to the fitted curve. This is statistically correct only if the experimental uncertainties of the independent and dependent variables are equal.

Other investigators attempt to treat uncertainties in an independent variable by reflecting it to a corresponding uncertainty in the dependent variable and then using an "appropriate weighting factor." This weighting factor is generally taken to be the inverse of the standard error of the dependent variable. When the fitted function in the region of a particular data point has a near-zero slope this procedure is equivalent to ignoring the experimental uncertainties in the independent variable. If, however, the fitted function has a significant slope in the neighborhood of the data point this procedure will transform small uncertainty in the independent variable into a large corresponding uncertainty in the dependent variable. This large dependent variable uncertainty will translate into such a small "appropriate weighting factor" that the net effect is that the data point will be

² The cross-correlation between parameters is a measure of the ability of one or more parameters to compensate for the variation of another parameter in a parameter estimation process. If the cross-correlation coefficient is plus or minus one, then for any reasonable value of one of the parameters, a set of the remaining parameters can be found which will yield the same numerical value for the statistical NORM. Consequently, a value of plus or minus unity indicates that unique values of the parameters cannot be estimated (1,5).

ignored by the least-squares procedure. Furthermore, if the fitted function has a significant curvature in the neighborhood of a particular data point it will introduce an asymmetrical non-Gaussian uncertainty in the dependent variable. This reflection procedure is incorrect as per Eq. [5].

The maximum likelihood approach, which minimizes the statistical NORM presented in Eq. [5], has been previously employed by other investigators [for example (6,7)]. Acton's application of the method is limited to only straight line data (6). Bard's book presents a different mathematical procedure to minimize the same statistical NORM (7). Bard's procedure determines the \bar{X}_i 's and the α 's simultaneously in a single iterative approach rather than the nested approach which was presented here. For each iteration Bard's approach requires the inversion of a sparse matrix of order equal to the number of independent variables times the number of data points plus the number of parameters being determined. As the number of data points increases the order of the matrix will increase and as a consequence the matrix will become difficult and time consuming to invert. In addition, as the order of the matrix increases its inversion will be more prone to problems caused by computational round-off

errors. The numerical method presented here may have some computational advantage over Bard's method as the number of data points increases even though the final result is mathematically equivalent.

It should be noted that even though the maximum likelihood method is considerably better than the more classical least-squares method for these two test examples, this is not a proof that this method is always better. It does, however, indicate that when Gaussian-distributed experimental uncertainties exist in both the independent and dependent variables this method of maximal likelihood estimation of the unknown parameters can prove to be very useful.

REFERENCES

1. Johnson, M. L., and Frasier, S. G. (1985) in *Methods in Enzymology*, Academic Press, New York, in press.
2. Caceci, M. S., and Cacheris, W. P. (1984) *Byte Magazine* **9**(5), 340.
3. Nelder, J. A., and Mead, R. (1965) *Comput. J.* **7**, 308.
4. Munson, P. J., and Rodbard, D. (1980) *Anal. Biochem.* **107**, 220.
5. Johnson, M. L. (1983) *Biophys. J.* **44**, 101.
6. Acton, F. S. (1959) *Analysis of Straight Line Data*, p. 129, Wiley, New York.
7. Bard, Y. (1974) *Nonlinear Parameter Estimation*, p. 67, Academic Press, New York.