# ANALYSIS OF DATA FROM THE ANALYTICAL ULTRACENTRIFUGE BY NONLINEAR LEAST-SQUARES TECHNIQUES

MICHAEL L. JOHNSON, *Diabetes Research and Training Center and Department of Pharmacology, University of Virginia School of Medicine, Charlottesville, Virginia 22908*

JOHN J. CORREIA AND DAVID A. YPHANTIS, *Biochemistry and Biophysics Section, Biological Sciences Group, University of Connecticut, Storrs, Connecticut 06268*

HERBERT R. HALVORSON, *Division of Biochemical Research, Henry Ford Hospital, Detroit, Michigan 48202*

ABSTRACT Least-squares analysis of experimental data from the analytical ultracentrifuge is discussed in detail, with particular attention to the use of interference optics in studying nonideal self-associating macromolecular systems. Several examples are given that describe the application of the technique, the expected precision of the results, and some of its limitations. A FORTRAN IV computer program is available from the authors.

## INTRODUCTION

Since the pioneering work of Svedberg and Pedersen (1), sedimentation equilibrium experiments have been analyzed by a linearization procedure consisting of a graph of the logarithm of the concentration as a function of the square of the radius. The slope of such a graph is a simple function of the molecular weight of the solute. This procedure assumes a single ideal nonassociating solute.

For more complex interacting systems, such as those exhibiting self-association or nonideality, this linearization procedure has been extended to measure average molecular weight moments such as number and weight averages as a function of concentration (2–7). Generally this is accomplished by estimating the rates of various derivatives obtained over limited ranges of concentration in the centrifuge cell. Values of equilibrium constants, monomer molecular weights, virial coefficients, etc. can then be evaluated from the behavior of these moments as a function of concentration and/or position. However, this method has two major pitfalls: the inherent difficulty and the associated (sometimes systematic) inaccuracies of numerically differentiating the data, and the difficulty in evaluating realistic confidence regions for parameters such as equilibrium constants arising from the use of the intermediate molecular weight moments and the consequent requirement of complicated weighting factors.

The exponential form of the concentration distribution has been used by several workers to evaluate equilibrium constants when the monomer molecular weights are known (8, 9). This approach has a serious drawback in that it requires knowing the molecular weight and partial specific volume of the monomer. However, the calculation is simple because it is a linear least-squares fit.

The nonlinear problem of an unknown monomer molecular weight has been discussed by several authors (8, 10–13). All of these approaches require estimation of the absolute concentration, instead of the relative concentration directly obtainable from interference optics. With one exception (13), these approaches also require that the solute be ideal.

The purpose of this publication is to describe and make available a FORTRAN IV program, NONLIN, which has been in use in our laboratories for some time (14–16). NONLIN performs a simultaneous nonlinear least-squares fit of one or more channels of ultracentrifuge data at different loading concentrations, radial positions, and possible angular velocities to a specific association and/or nonideality scheme. This nonlinear least-squares fit of the concentration distribution *directly* determines monomer molecular weights, virial coefficients, and association constants. Furthermore, the program only requires a relative concentration scale, such as is available from interference optics, instead of an absolute scale. Several examples of the use of this program appear in the literature (14–21). In addition to these examples, a similar program that uses the same algorithms has been employed for several other types of experimental data (22–24).

## FUNCTIONAL FORM

At sedimentation equilibrium the concentration distribution of the *i*th component of an ideal system is related to the effective reduced molecular weight, $\sigma_i$, and to the molecular weight, $M_i$, of this component by

$$\frac{\partial \ln c_{r,i}}{\partial(r^2/2)} = \sigma_i = \frac{M_i(1 - \bar{v}\rho)\omega^2}{RT} \tag{1}$$

where $c_{r,i}$ is the concentration of the *i*th component at a radius $r$, $R$ is the gas constant, $T$ is the absolute temperature, $\bar{v}$ is the partial specific volume, $\rho$ is the density and $\omega$ is the angular velocity. The concentration distribution of the *i*th species is obtained for constant $\sigma_i$ by integration of Eq. 1:

$$c_{r,i} = c_{o,i}\, e_i^{\sigma(r^2/2 - r_o^2/2)} \tag{2}$$

where $c_{o,i}$ is the concentration of the *i*th component at a radius $r_o$. For a monomer—*n*-mer association the total concentration at any radius, $c_{r,t}$, can be expressed in terms of the monomer concentration, $c_{r,1}$, and an association constant, $K_n$, by

$$c_{r,t} = c_{r,1} + K_n\,(C_{r,1})^n. \tag{3}$$

The equivalent of Eq. 1 for a nonideal system[1] describable by communal virial coefficients is

$$\frac{\partial \ln c_i}{\partial(r^2/2)} = \sigma_{i,a} = \frac{\sigma_i}{1 + \dfrac{\partial \ln \gamma_i}{\partial \ln c}} = \frac{\sigma_i}{1 + 2B_1c + 3B_2c^2 + \ldots} \tag{4}$$

[1] The variation of $\sigma_i$ with the density increment associated with the solute is formally the same, to first order, as a small positive contribution to the second virial coefficient, $B_1$ (32).

where $c$ is the total solute concentration ($\Sigma c_i$), the $\gamma_i$ are the activity coefficients of the $i$th species on the $c$ scale, and the virial coefficients of the system are denoted by $B_1, B_2 \ldots$.[2] In addition to the $\sigma_{i,a}$, the $c_{o,i}$, and to the communal $B_1, B_2 \ldots$, another parameter, $\delta c$, must be included for each channel used. This is an additive constant in the evaluation of the total concentration introduced because the Rayleigh interferometer usually measures relative concentration instead of absolute concentration. For the "meniscus depletion" technique this constant should be close to zero and is usually only a few micrometers of displacement on the photographic plate. However it must be included as

$$C_{\text{obs}} = C_{\text{true}} + \delta c \tag{5}$$

where the subscripts refer to the observed and true concentration (or fringe displacements). The corresponding concentration distribution for any association and/or nonideality scheme can be generated by appropriate combinations of Eqs. 2–5. The implicit assumptions of this approach are that no volume change occurs on association ($\bar{v}$ is a constant) and that the activity coefficients of the monomer, $\gamma_m$, and any $n$-mer, $\gamma_n$ are related by Eq. 6:

$$n \ln \gamma_m = \ln \gamma_n. \tag{6}$$

Under the assumption of Eq. 6 and of the mass action relation, the total concentration is given by:

$$C_{r,t} = \sum_{\ell=1}^{n} K_\ell \left\{ C_{0,1} \exp\left[ \sigma_1 (r^2/2 - r_o^2/2) - \sum_{k=1}^{m} \frac{k+1}{k} B_k C_{r,t}^k \right] \right\}^\ell \tag{7}$$

where $m$ is the total number of virial coefficients considered and $n$ is the number of associated species present. Accordingly we must fit the functional form, F:

$$Y_i \simeq F = \sum_{\ell=1}^{n} \left\{ C_{0,1} \exp\left[ (\ln K_\ell)/L(\ell) + \sigma_1 (r^2/2 - r_o^2/2) - \sum_{k=1}^{m} \frac{k+1}{k} B_k (F - \delta c)^k \right] \right\}^{L(\ell)} + \delta c \tag{8}$$

to the observations, $Y_i$. The criterion used is the familiar least-squares (minimization of the sum of the squares of the residuals).

This equation has been written for generality. It is usually assumed that $L(\ell) = \ell$, but for pressure-dependent associations we define $L(\ell) = \ell(1 - \bar{v}\rho)_\ell/(1 - \bar{v}\rho)_1$, as $\ell$ times the ratio of the buoyancy factor for the $\ell$-mer to that of the monomer and $K_\ell$ as the association constant at 1 atm pressure. The value of $K_1$ is taken as unity. This formulation accounts for the effects of a volume change on association. It is important to note that this implies that the apparent degree of association is not always an integer.

## NUMERICAL METHODS

The basic algorithm takes some function, $F$ (Eq. 8) and a series of data points, $Y_i$ and $X_i$, and determines a vector, $\boldsymbol{\alpha}$, of fitted parameters such that the sum of the squares of the residuals, the differences between

[2] The virial coefficients used in Eq. 4 are equal to the respective colligative virial coefficients multiplied by the monomer molecular weight. This form of Eq. 4 was chosen for mathematical simplicity, since the incorporation of the colligative virial coefficients in Eq. 4 would also require rotor speed, partial specific volume, and temperature to be included to calculate the monomer molecular weight.

the function and the data points, is a minimum. The numerical procedure used for this least-squares curve fitting is a modification of the basic Gauss-Newton procedure (25–27). This procedure is simply an algorithm which when given an initial guess for the vector $\alpha$ will find a better guess for $\alpha$. The procedure is then applied in an iterative fashion until the vector $\alpha$ does not change, within some specified tolerance.

The first step in any procedure of this type is to define the function, $F$. This function should predict the directly observed dependent variable, $Y_i$, rather than some function of $Y_i$ (such as $\ln Y_i$ used in the graphical procedure). The reason for this is that any nonlinear transformation made on $Y_i$ will distort the distribution of random experimental error (noise) which is always present on the data so that it may no longer be assumed to be Gaussian. With such distortion the usual least-squares criterion can no longer be applied. Consequently, we choose a function, $F$, to directly approximate the experimental data:

$$Y_i \simeq F_i = F(X_i, \alpha) \tag{9}$$

where $X_i$ is the corresponding independent variable.

The proper choice of fitting parameters, $\alpha$, is also important. For example, it is better to use the natural log of the equilibrium constant, $\ln K$, instead of the equilibrium constant. This forces the equilibrium constant to be positive at all times. In addition, since the free energy change is of greater interest, the error statistics are evaluated in free energy space by using $\ln K$ as one of the fitting parameters. In general, the curve fit should be done on the actual parameters which are of interest.

If we let the $i$ subscripts represent a particular data point and the $j$ subscripts represent a particular element of $\alpha$, we can then define a matrix $P$ whose elements are

$$P_{ij} = \frac{\partial F(X_i, \alpha)}{\partial \alpha_j} \tag{10}$$

and the vector of residuals, $\mathbf{Y}^*$, whose elements are

$$\mathbf{Y}_i^* = Y_i - F(X_i, \alpha). \tag{11}$$

The "correction vector," $\epsilon$, is then defined as

$$\epsilon = (P' P)^{-1} Py^* \tag{12}$$

where $P'$ is the transpose of $P$ and $(P' P)^{-1}$ is the inverse of the matrix $(P' P)$. Great care must be taken in finding the inverse of this matrix since this matrix is usually nearly singular. However, since the matrix is symmetric, the square root method of matrix inversion can be used (28). This square root method is exceptionally good for nearly singular matrices. The basic Gauss-Newton procedure then provides $\alpha^k$, the value of $\alpha$ for the $k$th iteration, as

$$\alpha^k = \alpha^{k-1} + \epsilon. \tag{13}$$

The procedure is repeated until $\alpha$ does not change within some specified tolerance, usually a fractional change of one part in a million.

The Gauss-Newton procedure was modified so that instead of using $\epsilon$ to provide both direction and distance it is used only for the direction and a search is made to find the distance which gives a minimum variance. This search is performed by either multiplying or dividing the magnitude of $\alpha$ by two until two distances are found: one whose corresponding variance is less than the variance of the previous iteration and one whose corresponding variance is larger than the previous iteration. The value from'the previous iteration and these points are then fit to a quadratic polynomial. The resulting polynomial is then differentiated to find the distance corresponding to the minimum variance. Occasionally one of the points used in the search will have a lower variance than the predicted minimum, and in this case the lowest point is taken. Such a search procedure forces the variance to decrease with each iteration. This greatly improves the convergence properties of the Gauss-Newton procedure thus relaxing the requirements on the initial guesses. This modification is a generalization of a procedure suggested by Box (25) and Hartley (26).

A complication arises for cases where the function, $F$, is a function of itself (e.g., for nonideal sedimentation equilibrium, Eqs. 7 and 8). This can be resolved by defining a new function, $G$, as the difference between the calculated concentration and the function $F$ evaluated at that calculated concentration. The value of the function $F$ can then be evaluated for each data point iteratively by Newton's method as

$$F^{q+1} = F^{q} - \partial G/\partial F^{q} \tag{14}$$

where $q$ refers to the iteration number. The iterative process is continued until $F^{q+1}$ becomes indistinguishable from $F^{q}$; i.e., $G$ approaches zero.

The reported confidence limits are calculated by searching the variance space for an $F$-statistic corresponding to approximately a 65% confidence probability (22, 23). The 65% confidence region for a Gaussian distribution is the mean plus or minus roughly one standard deviation. This search of the variance space is performed in two ways: In the first way each of the elements of $\alpha$ is varied independently. In the other way the direction of the search is defined in terms of the $F$ statistic (variance ratio) as the axis of the multidimensional hyperellipsoid defined by the solutions, $\Omega$, of the following matrix equation (25, 27).

$$(\alpha - \Omega)\, PP'\, (\alpha - \Omega) < n\, s^{2}\, F \,\text{statistic}, \tag{15}$$

where $n$ is the number of parameters and $s^2$ is the variance of the minimum. Confidence intervals evaluated by this procedure correspond to approximately 1 SD, but because of the correlation between successive data points and between parameters these confidence limits are only estimates of the true value. In general such confidence limits will be asymmetrical and are thus reported as a range of values instead of a single value. These confidence limits only reflect the precision of the fit of the experimental data to the model and do not necessarily indicate the accuracy of the determined parameters. The evaluation of the confidence region does not include possible effects of systematic errors in the data. Likewise the use of an incorrect association and/or nonideality model can lead to utter nonsense (examples of this are given later).

The cross-correlation between fitting parameters, $CC_{ij}$, is evaluated from the elements of the inverse of the $P'P$ matrix at the solution,

$$C_{ij} = (P'P)^{-1}_{ij}/[(P'P)^{-1}_{ii}(P'P)^{-1}_{jj}]^{1/2}. \tag{16}$$

Goodness of fit can be determined by two commonly used criteria: first, the variance must be approximately the same as the experimental noise level, e.g., a few micrometers of displacement on the photographic plate; second, the residuals must appear to be random as a function of either concentration or radius. The latter criterion is usually applied qualitatively rather than quantitatively. The program does, however, quantitate this second criterion by estimating the nonrandomness of the residuals.

To evaluate the precision and range of applicability of these algorithms, simulated data with pseudo-random noise was used. For the purpose of testing the program simulated data are preferable because (*a*) the "correct" answers and model are known, (*b*) the amount of added random noise is also known, and (*c*) the same data can be analyzed with different sets of random noise. Several examples already exist in the literature where these algorithms have been used on real experimental systems (14–21, 23).

The simulation of data involved in iterative process to choose monomer concentrations at some reference points, consistent with conservation of mass in an ultracentrifuge cell with specific bounding radii and loading concentrations. The radii were chosen to simulate the six-channel centerpieces available for the Beckman Model E ultracentrifuge (29, 30) (Beckman Instruments, Palo Alto, Calif.). Initial loading concentrations were chosen so that the concentrations of the solute in the three sample cells were in a ratio of 9:3:1 (inside to outside radius) and such that the middle cell would have the maximum observable fringe gradient (15 mm/cm$^2$) at its base. In addition, the data from all the cells were truncated when the concentration gradient exceeded a realistic experimental value, 15 mm/cm$^2$. A total of approximately 80–90 equally spaced points were simulated for each of the three pairs of channels of the centerpiece. Data simulated in this manner are consequently a good approximation of what can be obtained for real experimental systems with the analytical ultracentrifuge.

In an attempt to further simulate real experimental data, Gaussian distributed random noise of some specified amplitude, usually 3 $\mu$m on the photographic plate, was added to each of the simulated sets of data. These Gaussian distributed random numbers were calculated as 6 minus the sum of 12 random numbers evenly distributed between 0 and 1. The resulting numbers have a mean of zero and an SD of 1 and consequently can be scaled to give any desired mean and standard deviation. The evenly distributed random numbers were calculated with functions RANDU, (IBM scientific subroutine package; IBM Corp., White Plains, N.Y.) or RANF (Control Data Corp., Minneapolis, Minn.), depending on the computer used.

## TESTS OF METHODS

The description of the function and usefulness of this program consists of a series of examples. These examples include the analysis of a monomer $n$-mer association (a relatively simple case) and a nonideal monomer dimer association (a particularly difficult problem). In addition, some examples illustrate possible pitfalls in analyzing data of this type by any method. First, however, we demonstrate that our method of finding the confidence interval, the use of Eq. 15, is reasonable for at least one functional form.

### *Confidence Intervals*

To test the algorithm for evaluating the confidence intervals a linear system must be used. The use of a linear system, in this case a quadratic polynomial, allows independent calculation of the confidence intervals by standard statistical techniques. In this manner we have an estimate of the true confidence intervals to compare with that generated by the algorithm which we have used.

To this end, 25 evenly spaced data points were calculated for a quadratic polynomial:

$$\gamma = B_0 + B_1X_i + B_2X_i^2 \qquad 0 < X_i < 1$$

where $B_0 = 4$, $B_1 = 3$ and $B_2 = 2$. Gaussian distributed simulated random noise was then superimposed on these data so that the standard deviation of this noise was 0.02.

These data were then "analyzed" by the algorithms presented in the Numerical Methods section yielding among other things the confidence interval for each of the parameters. In addition, the standard error of each parameter was evaluated by the standard statistical method, taking the variance of each of the parameters as equal to the variance of the data points times the corresponding diagonal element of the $(P'P)^{-1}$ matrix. Table I presents a comparison of these standard deviations and of the confidence intervals for these calculations. The example given in Table I shows that for this case the values of the confidence interval as predicted by Eq. 15 are approximately a factor of 2 larger than those predicted by the more standard techniques. This discrepancy probably stems from a difference in the assumptions involved in evaluating these two statistical measures. The usual method of evaluating the standard error assumes that the parameters are independent. It is, however, easy to demonstrate that this is not the case. Table II gives the cross-correlation matrix for the curve fit presented in Table I. The correlation matrix implicitly describes the shape (but not the size) of the confidence region. If all off-diagonal elements were zero, this region would be

TABLE I

A COMPARISON OF STANDARD ERRORS AND CONFIDENCE INTERVALS FOR EACH OF THE PARAMETERS WHEN FITTING TO A QUADRATIC POLYNOMIAL

| Parameter | Value[*] | Confidence interval[§] | Standard error[¶] |
|---|---|---|---|
| $B_0$ | 4.000 | (3.970, 4.030) | ±0.015 |
| $B_1$ | 2.980 | (2.837, 3.124) | ±0.066 |
| $B_2$ | 2.025 | (1.896, 2.154) | ±0.062 |

[*] Best least-square value.

[§] As determined by Eq. 15.

[¶] As determined by standard statistical methods (see text).

radially symmetric about the estimated values for the parameters: a hypersphere. In the presence of correlation, the eigenvalues of the correlation matrix describe the lengths of the principal semi-axes of the hyperellipsoid and the corresponding eigenvectors define their orientation (25). These are the directions used in describing the confidence region. For the matrix in Table II, the eigenvalues are 2.745, 0.246, and 0.008. Hence along an unique linear combination of parameters, the confidence region appears some 2.75 times larger than predicted by simple linear theory. In other directions, however, the confidence region is more tightly defined. With the high cross-correlation shown in Table II it is not surprising that the standard statistical method predicts a value that is a factor of 2 less than an alternative method that includes cross-correlation as one of its assumptions. Consequently, Eq. 15 seems to predict a reasonable value for the confidence interval, at least for this example.

A further word is in order about confidence intervals. The expected confidence intervals with nonlinear functions are rarely symmetrical and thus is is impossible to express them as plus or minus a single value. An example of an asymmetrical confidence interval is the confidence interval of an equilibrium constant, or ln $K$, when the experimental conditions are such that the equilibrium is almost completely in either direction. In this circumstance the confidence region will be very asymmetrical with one of the limits being either plus or minus infinity. Our method of evaluating the confidence interval allows for asymmetric confidence intervals and will indeed predict a confidence limit of plus or minus infinity as required by the data!

### *Ideal Self-association*

The example that was chosen to illustrate the usefulness of this method of analysis is a monomer-tetramer equilibrium. For each of the cases investigated data were calculated for the given equilibrium in an attempt to simulate an experiment using the available six-channel

TABLE II

THE CORRELATION MATRIX[*] FOR THE ANALYSIS PRESENTED IN TABLE I.

| Parameter | $B_0$ | $B_1$ | $B_2$ |
|---|---|---|---|
| $B_0$ | 1.000 | −0.879 | 0.765 |
| $B_1$ | −0.879 | 1.000 | −0.970 |
| $B_2$ | 0.765 | −0.970 | 1.000 |

[*] Calculated according to Eq. 16.

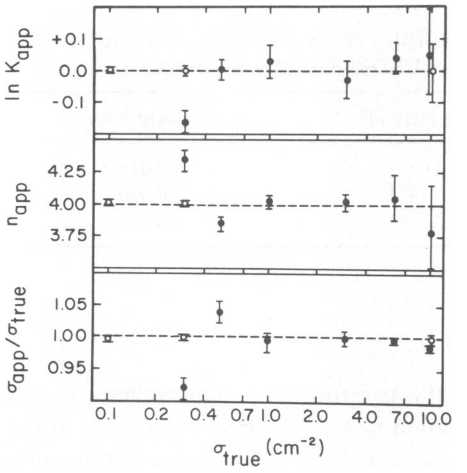

FIGURE 1

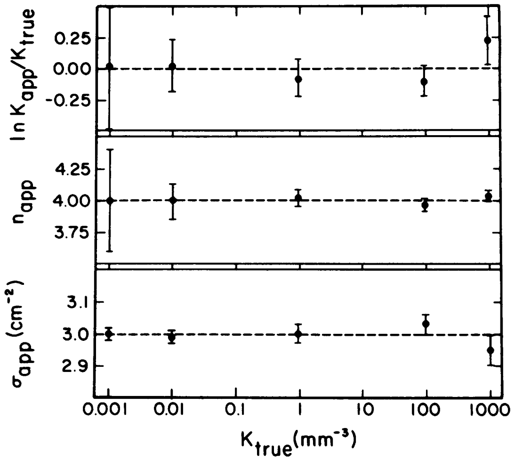

FIGURE 2

FIGURE 1 The effect of a variation of the reduced molecular weight $\sigma$ on the ability to determine $\sigma$, $\ln K$, and $n$ for a monomer-tetramer system. See text for details of data simulation and analysis.

FIGURE 2 The effect of a variation of the equilibrium constant on the ability to determine $\sigma$, $\ln K$, and $n$. See text for details of data simulation and analysis.

centerpiece. Normally distributed pseudo-random noise was added to the data at a realistic level for careful measurement (0.003 mm SD). Unless otherwise stated, the equilibrium simulated is a monomer-tetramer with an equilibrium constant[3] of 1.0 $\text{mm}^{-3}$ and a reduced molecular weight, $\sigma$, of 3.0 $\text{cm}^{-2}$.

The effect of a variation of the reduced molecular weight, $\sigma$, on the ability to estimate $\sigma$, the natural logarithm of the equilibrium constant, $\ln K$, and the degree of polymerization, $n$, is shown in Fig. 1. The open circles below $\sigma$ of $\sim$ 0.5 $\text{cm}^{-2}$ correspond to determinations where the base-line concentration level is known ($\delta c = 0$); under these conditions this could be obtained through a complementary synthetic boundary experiment, and the assumption of mass conservation. The open circles at $\sigma = 10\ \text{cm}^{-2}$ correspond to an analysis where the degree of association is known ($n = 4$). Between values of $\sigma$ of $\sim$0.7–7 $\text{cm}^{-2}$, in this example, this method can simultaneously evaluate monomer molecular weights to within a few percent, and free energies of association to within 60 cal/mol without prior knowledge of the degree of association.

A similar series of calculations is given in Fig. 2, where the equilibrium constant is varied over six orders of magnitude: $\sigma = 3\ \text{cm}^{-2}$; $n = 4$. With the exception of values of the equilibrium constant of 0.001 $\text{mm}^{-3}$ it can be seen that the actual value of the equilibrium constant has little effect on the estimation of itself, the degree of association or the reduced molecular weight. The inability to accurately determine these parameters at values of $K <$ 0.001 $\text{mm}^{-3}$ is not surprising since then the weight fraction of tetramer never exceeds 5% in any of the cells and is only 0.1% at the base of the least concentrated cell.

[3] The unit of concentration used in these simulations is millimeters of displacement of the interference pattern on the photographic plate. In a typical model E ultracentrifuge using 12-mm centerpieces this corresponds to ~1 mg/ml.

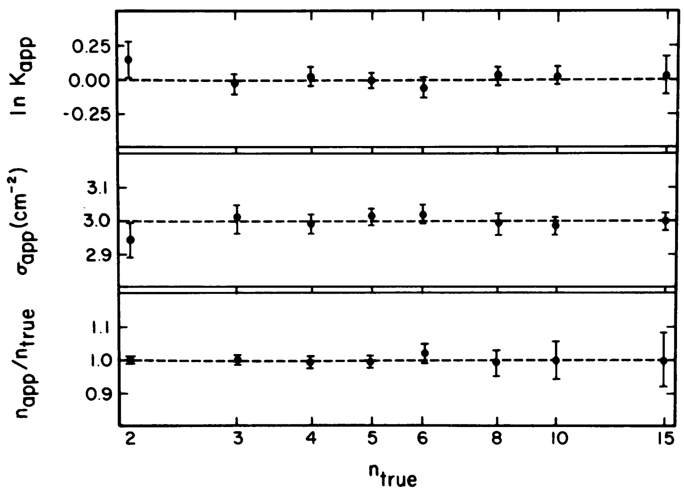

FIGURE 3 The effect of a variation of the degree of polymerization on the ability to determine $\sigma$, ln $K$, and $n$. See text for details of data simulation and analysis.

A similar variation of the degree of polymerization ($n$) is shown in Fig. 3. It is obvious that degrees of polymerization as high as 15 have little effect on the analysis.

### *Nonideal Self-association*

It has been demonstrated analytically that the variation in molecular weight that results from self-association can, under some conditions, be expressed as a series expansion in concentration similar to Eq. 4 (31). It was further demonstrated that for a monomer–$n$-mer association the second term is proportional to the concentration to the $n - 1$ power and has a negative sign. Consequently, a nonideal monomer-dimer association should be and is, exceptionally difficult to resolve since the terms of the respective expansions tend to cancel each other. This is the example we have chosen to demonstrate the use of the method.

Synthetic data were generated as before with various values of the virial coefficient, $B_1$. In these examples $K$ was taken to be 1.0 mm$^{-1}$ and $\sigma$ to be 5.0 cm$^{-2}$. Fig. 4 shows the results of the analysis of these data at several values of $B_1$. It appears that for a single experiment with three solution/solvent pairs the lower limit for the measurement of $B_1$ (simultaneously with $K_2$) is 0.004 mm$^{-1}$. To enable the reader to easily appreciate the meaning of various values of $B_1$ we have presented in Fig. 5 the apparent weight-average value of the reduced molecular weight of the solution as a function of concentration for three values of $B_1$. Comparing the lower limit of measurement, 0.004 mm$^{-1}$, for $B_1$ with Fig. 5 indicates that this is actually very little nonideality. The inset in Fig. 5 is to illustrate that the lower limit of measurement corresponds to a change in the shape of the graph of apparent weight-average molecular weight against log concentration. It is this change in shape that will enable $B_1$ and $K_2$ to be resolved. The upper limit appears to be between 0.3 and 1.0 mm$^{-1}$, which is sufficient nonideality that the association is very difficult to observe (see Fig. 5).

Another interesting phenomenon is illustrated in Fig. 4. The points at $B_1$ values of 0.003 and 0.001 mm$^{-1}$ show a high correlation between the values of ln $K$ and $\sigma$. It should be pointed out, however, that under these conditions, the weight fraction dimer does not exceed

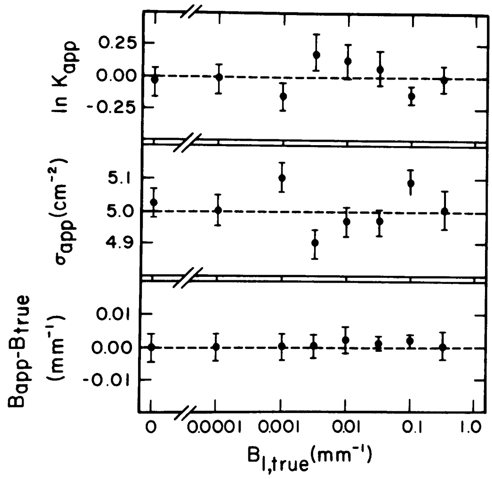

FIGURE 4

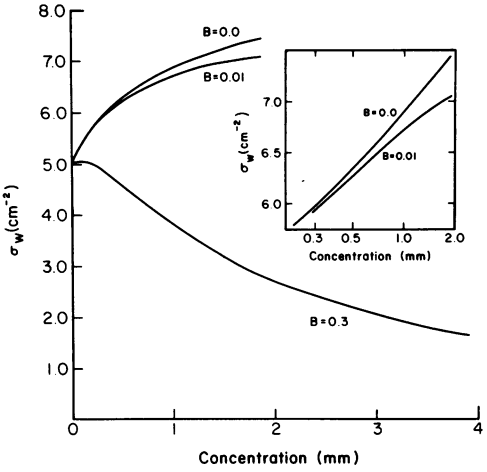

FIGURE 5

FIGURE 4 The effect of a variation of the amount of nonideality on the ability to determine $\sigma$, ln $K$, and $B_1$ for a nonideal monomer-dimer self-association: $\sigma = 5\ \text{cm}^{-2}$, $K = 1\ \text{mm}^{-1}$, $n = 2$. See text for the details of data simulation and analysis.

FIGURE 5 Weight-average-reduced molecular weight as a function of concentration for some of the nonideal examples presented in Fig. 4.

50% anywhere in any of the cells. It is consequently not surprising that the cross-correlation coefficient between the parameters is −0.978 and that this is reflected in the analysis. What is surprising is that in a situation such as this the molecular weight can be determined to within 2% and the free energy of association to within 200 cal/mol.

### *Incorrect Models*

A second nonideal example is worth nothing. In this example data were generated to simulate an ideal monomer-dimer association with a maximum weight fraction dimer of 23% at the base of the most concentrated cell. These data were then analyzed as an ideal monomer–$n$-mer association and as a nonideal single component. A comparison of the resulting parameter values is given in Table III. There are several noteworthy points in this table. First, the choice of an incorrect model does not adversely affect the ability to determine the molecular weight, for this example. This may be useful, but is not guaranteed to be true in every case. Second, the second virial coefficient, $B_1$, is negative. The choice of an incorrect model may well be the source of some of the negative values of $B_1$ reported in the literature. Third, the analysis by the incorrect model picked values of the base line, $\delta c$, which are incorrect by ~1/100 of a fringe on the photographic plate. This emphasizes the need of a very accurate determination of this parameter whenever possible. Fourth, the incorrect model yielded a variance which was 25% higher than the correct model. Since the critical $F$ statistic for a 65% confidence interval is <1.05, the correct model can be ruled out on this basis. It should be noted, however, that the incorrect model appears to fit the data by all other criteria which can be applied including randomness of the residuals. Consequently, if there is any possibility of systematic or nonrandom noise in the data these two models can probably not be distinguished.

TABLE III

ANALYSIS OF DATA GENERATED TO SIMULATE A MONOMER-DIMER ASSOCIATION WITH THE CORRECT AND INCORRECT NONIDEAL NONASSOCIATING MODEL*

| | Analyzed as | |
|---|---|---|
| | monomer-dimer | nonideal monomer |
| $\sigma$ | 3.010 (2.986, 3.035) | 3.098 (3.079, 3.117) |
| $n$ | 2.025 (1.971, 2.079) | |
| $\ln K_2$ | −2.377 (−2.556, −2.204) | |
| $B_1$ | | −0.024 (−0.025, −0.023) |
| $\delta c_1$ | 0.0006 (−0.0005, 0.0018) | 0.0028 (0.0012, 0.0043) |
| $\delta c_2$ | 0.0004 (−0.0007, 0.0015) | 0.0023 (0.0009, 0.0036) |
| $\delta c_3$ | 0.0001 (−0.0011, 0.0013) | 0.0016 (0.0002, 0.0031) |
| Variance | $8.76 \times 10^{-6}$ | $11.17 \times 10^{-6}$ |

*$\sigma = 3.0\ \text{cm}^{-2}$; $K_2 = 0.2\ \text{mm}^{-1}$; $n = 2$.

Another common mistake in the literature is the neglect of small amounts of nonideality. Every macromolecule can potentially exhibit nonideality because of excluded volume and charge repulsion. In a carefully designed experiment these effects often can be diminished, but cannot be eliminated. The reader is referred to our experiments on *Limulus* hemocyanin performed in 0.45 M KCl (16). Even with nearly half molar salt we were able to measure the nonideality and simultaneously account for the nonideality as a combination of charge effect (predominant) and excluded volume.

In Fig. 6 we present the apparent values of the reduced molecular weight, $\sigma$, and the logarithm of the equilibrium constant, ln $K$, when a nonideal monomer-dimer association is analyzed as an ideal monomer-dimer association. Comparison with Fig. 5 shows that a value of 0.01 $\text{mm}^{-1}$ for $B_1$ does not appear to be markedly nonideal, yet neglect of this small amount of nonideality does indeed cause significant errors.

To demonstrate that similar problems exist for ideal systems we present in Table IV an analysis of a monomer-trimer association in terms of the correct monomer-trimer model and in terms of an incorrect monomer-dimer-tetramer model. In this incorrectly analyzed case, the

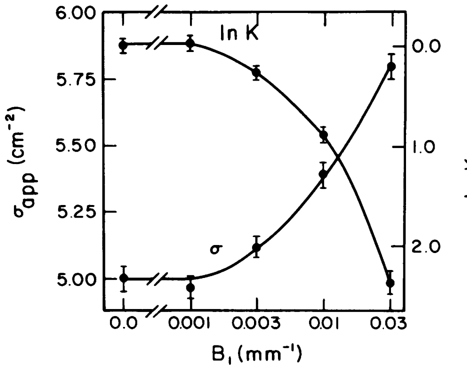

FIGURE 6 Errors introduced into the evaluation of $\sigma$ and ln $K$ by neglecting various amounts of nonideality. Data were generated for a monomer-dimer association with varying amounts of nonideality and analyzed assuming ideality: $\sigma = 5.0\ \text{cm}^{-2}$; $K = 1.0\ \text{mm}^{-1}$. See text for details.

TABLE IV
ANALYSIS OF DATA GENERATED TO SIMULATE A MONOMER-TRIMER MODEL BY THE CORRECT AND AN INCORRECT MONOMER-DIMER-TETRAMER MODEL*

| | Analyzed as | |
|---|---|---|
| | monomer-trimer | monomer-dimer-tetramer |
| $\sigma$ | 3.014 (2.984, 3.043) | 2.610 (2.571, 2.648) |
| $\ln K_3$ | −0.029 (−0.089, +0.034) | |
| $\ln K_2$ | | 0.422 (0.322, 0.525) |
| $\ln K_4$ | | 1.180 (1.033, 1.330) |
| $\delta c_1$ | +0.0002 (−0.0009, 0.0012) | −0.0024 (−0.0036, −0.0012) |
| $\delta c_2$ | +0.0002 (−0.0009, 0.0012) | −0.0022 (−0.0034, −0.0011) |
| $\delta c_3$ | +0.0000 (−0.0010, 0.0011) | −0.0019 (−0.0031, −0.0007) |
| Variance | $7.78 \times 10^{-6}$ | $9.51 \times 10^{-6}$ |

$\sigma = 3.0\ \text{cm}^{-2}$; $K_3 = 1.0\ \text{mm}^{-2}$; $n = 3$.

monomer molecular weight is significantly in error, and the equilibrium constants are totally meaningless. As mentioned in the discussion of Table III, the base-line error is small but significant; again the variance is significantly higher for the incorrect model. However, as previously mentioned, the lower variance is only a weak argument in favor of one model over the other if the possibility of systematic experimental errors exists.

It should be noted that these limitations and consequences imposed by the use of an incorrect model, i.e., neglecting nonideality etc., are not a problem or limitation specific to this method of analysis, or to this computer program, and to a large extent are not limitations imposed by the use of the analytical ultracentrifuge. These problems are in fact a consequence of the similarities of the functional forms of the conservation of mass equations for the different molecular interactions and the physical limits of solute concentration and detection imposed by any experiments on macromolecules.

## CONCLUSIONS

In this report we have presented and demonstrated the usefulness of nonlinear least-squares techniques as applied to the analytical ultracentrifuge. The examples used to test both the method and our computer program were specifically chosen to demonstrate the flexibility and precision of the approach for both ideal and nonideal systems. In addition, several examples were chosen which point out pitfalls of an incorrect choice of model.

We wish to emphasize that the primary usefulness of this method of analysis is to test particular models against experimental data and determine the corresponding parameter values and confidence intervals. It can be used to rule out models that obviously do not fit the experimental data. However this approach cannot easily be used to distinguish between different models that fit the data with approximately the same precision. Consequently, this method should be used in conjunction with some of the model independent methods that have been developed (2–4, 7).

We have used this general method of analysis for approximately 10 years in the analysis of data from the analytical ultracentrifuge and from other biochemical techniques. The reader is referred to the literature for a discussion of these experimental systems (15–24) and to recent studies from other workers (33, 34).

Should the reader desire to apply these methods to the analytical ultracentrifuge, two versions of a FORTRAN IV program are available from the authors: one for an IBM 370 and the other for a Control Data Corp. Cyber 720. A third version is also available that is not specific to the analytical ultracentrifuge, but requires a user-specified equation.

Special thanks are extended to Dr. G.K. Ackers; a large portion of this work was performed indirectly while Drs. Johnson and Halvorson were postdoctoral fellows in this laboratory.

This work was supported by the National Science Foundation, GB-13790, BMS 71-01299, PCM 76-21847, and BMS 74-24507 and by the U.S. Public Health Service grants GM-14493, AM 22125, and GM 28928. In addition Dr. Johnson was supported by a U.S. Public Health Service predoctoral training grant, GM-00317, and a postdoctoral Fellowship, AM 03201.

*Received for publication 24 February 1981 and in revised form 7 July 1981.*

## REFERENCES

1. Svedberg, T., and K. O. Pedersen. 1940. The Ultracentrifuge. Oxford University Press, London. 467 pp.
2. Roark, D., and D. A. Yphantis. 1969. Studies of self-associating systems by equilibrium ultracentrifugation. *Ann. N. Y. Acad. Sci.* 164:245–278.
3. Wan, P. J., and E. T. Adams, Jr. 1976. Molecular weights and molecular weight distributions from ultracentrifugation of nonideal solutions. *Biophys. Chem.* 5:207–241.
4. Tang, L. H., D. R. Powell, B. M. Escott, and E. T. Adams, Jr. 1977. Analysis of various indefinite self-associations. *Biophys. Chem.* 7:121–139.
5. Kim, H., R. C. Deonier, and J. W. Williams. 1977. The investigation of self-association reactions by equilibrium ultracentrifugation. *Chem. Rev.* 11:659–690.
6. Lewis, M. S., and G. D. Knott. 1976. Simulation studies of self-associating systems: discrimination between specific and isodesmic associations. *Biophys. Chem.* 5:171–183.
7. Stafford, W. F., III. 1980. Graphical analysis of nonideal monomer *N*-mer, isodesmic, and type II indefinite self-associating systems by equilibrium ultracentrifugation. *Biophys. J.* 29:149–166.
8. Haschemeyer, R. H., and W. F. Bowers. 1970. Exponential analysis of concentration or concentration difference data for discrete molecular weight distributions in sedimentation equilibrium. *Biochemistry.* 9:435-445.
9. Holladay, L. A., and A. J. Sophianopoulos. 1972. Nonideal associating systems: documentation of a new method for determining the parameters from sedimentation equilibrium data. *J. Biol. Chem.* 247:427–439.
10. Rosenthal, A. 1971. Analysis of polydisperse systems at sedimentation equilibrium. I. Simple solvent systems. *Macromolecules.* 4:35–42.
11. Williams, R.C. 1971. Analyses of some associating systems by equilibrium ultracentrifugation. *162nd ACS.* BIOL 277 *Abstr.*
12. Kar, E. G., and K. C. Anne. 1974. Analysis of sedimentation equilibrium data. *Anal. Biochem.* 62:1–18.
13. Milthorpe, B. K., P. D. Jeffrey, and L. W. Nichol. 1975. The direct analysis of sedimentation equilibrium results obtained with polymerizing systems. *Biophys. Chem.* 3:169–176.
14. Johnson, M. L., and D. A. Yphantis. 1971. Nonlinear least squares of sedimentation equilibrium data. *162nd ACS* BIOL 278 *Abstr.*
15. Johnson, M. L. 1973. Subunit structure of *Limulus* hemocyanin. Ph.D. Thesis., University of Connecticut, Storrs, Conn.
16. Johnson, M. L., and D. A. Yphantis. 1978. Subunit structure and heterogeneity of *Limulus polyphemus* hemocyanin. *Biochemistry.* 17:1448–1455.
17. Szuchet, S. 1976. Equilibrium centrifugation of proteins in acidic solutions. *Arch. Biochem. Biophys.* 177:437–460.
18. Szuchet, S., and D. A. Yphantis. 1976. Equilibrium centrifugation of proteins in acidic solutions. *Arch. Biochem. Biophys.* 173:495–516.
19. Yphantis, D. A., and M. L. Johnson. 1971. Heterogeneity in self-associating systems. *Biophys. Soc. Abstr.* WPM-H18.

20. Yphantis, D. A., M. L. Johnson, and G. M. Wu. 1972. Effects and detection of heterogeneity in self-associating systems. *Abstr. 163rd ACS* AGFD 2.

21. Yphantis, D. A., J. J. Correia, M. L. Johnson, and G. M. Wu. 1979. Detection of heterogeneity in self-associating systems. *In* Physical Aspects of Protein Interactions. N. Catsimpoolas, editor. Elsevier/North Holland, Amsterdam. 275–303.

22. Ackers, G. K., M. L. Johnson, F. C. Mills, H. R. Halvorson, and S. Shapiro. 1975. The linkage between oxygenation and subunit dissociation in human hemoglobin: consequences for the analysis of oxygenation curves. *Biochemistry* 14:5128–5134.

23. Johnson, M. L., H. R. Halvorson, and G. K. Ackers. 1976. Oxygenation-linked subunit interactions in human hemoglobins: analysis of the linkage functions for constitutent energy terms. *Biochemistry.* 15:5363–5371.

24. Turner, B. W., D. W. Pettigrew, and G. K. Ackers. 1981. Measurement and analysis of ligand-linked subunit association equilibria in human hemoglobins. *Methods Enzymol.* 76: In press.

25. Box, G. E. P. 1960. Fitting empirical data. *Ann. N. Y. Acad. Sci.* 86:792–816.

26. Hartley, H. O. 1961. The modified Gauss-Newton method for the fitting of nonlinear regression functions by least squares. *Technometrics.* 3:269–280.

27. Magar, M. 1972. Data analysis in biochemistry and biophysics. Academic Press, Inc., New York. 149–243.

28. Faddeeva, V. N. 1959. Computational Methods of Linear Algebra. Dover, New York. 81.

29. Yphantis, D. A. 1964. Equilibrium ultracentrifugation of dilute solutions. *Biochemistry.* 3:297–317.

30. Ansevin, A. T., D. E. Roark, and D. A. Yphantis. 1970. Improved ultracentrifuge cells for high-speed sedimentation equilibrium studies with interference optics. *Anal. Biochem.* 34:237–261.

31. Stafford, W., and D. A. Yphantis. 1972. Virial expansions for ideal self-associating systems. *Biophys. J.* 12:1359–1365.

32. Fujita, H. 1962. Mathematical theory of sedimentation analysis. Academic Press, Inc., New York.

33. Gladner, J. A., M. S. Lewis, and S. I. Chung. 1981. Molecular properties of lamprey fibrinogen. *J. Biol. Chem.* 256:1772–1781.

34. Davies, P. J. A., D. Wallach, M. Willingham, I. Pastan, and M. S. Lewis. 1980. Self-association of chicken gizzard filamin and heavy merofilamin. *Biochemistry.* 19:1366–1372.