Computer-Assisted Analysis of Light-Scattering Data by Linear and Nonlinear Least-Squares Methods

Donald B. Siano* and Jon Applequist*

Department of Biochemistry and Biophysics, Iowa State University, Ames, Iowa 50010.
Received June 4, 1975

ABSTRACT: A linear least-squares method is described for fitting conventional light-scattering data (Kc/R_θ) on macromolecular solutions to a polynomial in the concentration c and scattering angle function $x = \sin^2(\theta/2)$. Terms through c^2 and x^3 and a cross term cx are included on the basis of goodness of fit tests showing that such terms are significant in practice. A least-squares program LS3 evaluates the molecular parameters and their standard deviations, including shape and polydispersity parameters which are determined by the coefficients of x^2 and x^3. The method is applied to real data on poly(LysZ) in dimethylformamide and to simulated data for particles of known molecular parameters to test the range of applicability. Systematic errors which become significant for large particles because of the polynomial approximation are evaluated for rods and random coils. For treatment of large particles a nonlinear least-squares program NLN is described which finds an optimum fit of observed data to scattering functions calculated exactly for various trial shapes. The method is shown to distinguish the shape and to give accurate molecular parameters for simulated scattering data on rods.

Computer analysis of light-scattering data has been the subject of several recent studies[1-4] and one review[5] which describe methods for obtaining greater precision and accuracy in the desired molecular parameters with greater ease than is obtained by the traditional graphical methods. These studies have made use of linear least-squares methods to fit the scattering function Kc/R_θ to a polynomial of the form

$$Kc/R_\theta = \sum_{k=0}^{m} \sum_{l=0}^{n} a_{kl} c^k x^l \qquad (1)$$

where K is an optical constant, c is the weight concentration, R_θ is the Rayleigh ratio at scattering angle θ, and $x = \sin^2(\theta/2)$. The coefficients a_{kl} are found in the fitting process, and are related to the molecular parameters. The various treatments have differed primarily in the degree assigned to the polynomial, expressed in (1) by m and n. Thus Evans et al.[1] applied the case $m = 2$, $n = 1$, and omitted the cross terms on theoretical grounds. Bryce[2] treated the case $m = n = 2$, and omitted all cross terms except c^2x, citing theoretical grounds for inclusion of the latter term. Miller and Stepto[3] treated the case $m = n = 2$ more generally, using the fit to experimental data to determine which cross terms should be included. These authors found that the need for cross terms varies from one set of experimental data to another. Mijnlieff and Coumou[4] used a fitting procedure in which data extrapolated to zero concentration are fitted to a polynomial in x, and found that terms through x^3 are needed to fit certain data on polymer solutions.

Attempts in this laboratory to apply similar procedures to light-scattering data on synthetic polypeptides have led to the conclusion that none of the polynomials recommended by previous authors is entirely satisfactory for these systems due to the omission of cross terms or of terms higher than x^2. The purpose of this paper is to illustrate the use of statistical criteria to select the optimum terms from a polynomial of the form in (1) to fit real data. We investigate the magnitudes of both random and systematic errors in molecular parameters arising from the use of this fitting process. Simulated data for molecules of known size and shape and with known levels of random noise are used as test cases. In addition, we suggest here a nonlinear least-squares method that is sometimes helpful both in eliminating the systematic errors present in the polynomial fit and in determining the particle shape and polydispersity

Selection of Polynomial Terms

Zimm[6] and Bryce[2] have obtained the expression

$$Kc/R_\theta = 1/MP(\theta) + 2A_2c + \{3A_3Q_1(\theta) + 4A_2^2MP(\theta)[(1 - P(\theta))]\}c^2 + \ldots \qquad (2)$$

where M is the molecular weight, $P(\theta)$ is the particle scattering factor, $Q_1(\theta)$ is a scattering factor due to intermolecular interference, and A_2 and A_3 are the second and third virial coefficients, respectively. On the other hand, Flory and Bueche[7] obtained for Gaussian chains

$$Kc/R_\theta = 1/MP(\theta) + 2A_2Q_2(\theta)c + \{3A_3Q_3(\theta) + 4A_2^2M[P(\theta)Q_2^2(\theta) - Q_3(\theta)]\}c^2 + \ldots \qquad (3)$$

where $Q_2(\theta)$ and $Q_3(\theta)$ are other intermolecular scattering factors. A comparison of (2) and (3) shows that the nature of terms involving both c and x is different in the two. Particularly, the term cx is absent in (2) but present in (3), if $Q_2(\theta)$ is regarded as a power series in x. This distinction arises from the inclusion of higher order intermolecular contacts in the treatment of Flory and Bueche, as opposed to the single-contact approximation in Zimm's treatment. Since it is not usually known in advance which level of approximation is suitable for a given system, we adopt the view here that the need for cross terms, as well as other terms, should be determined by the goodness of fit of various polynomials to experimental data.

The selection of the optimum set of independent variables to use in a regression program is an often encountered, much discussed problem in statistics, and various approaches are possible. A paramount consideration, however, is the choice of the "basic" set of terms containing all of the terms that are believed to be possibly necessary for the explanation of the data. This set of terms must be obtained from one's knowledge of the origin of the data, its physical meaning, and past experience with less refined (e.g., graphical) methods in the analysis of similar data. The fact that our polynomial set is a Taylor series expansion of (2) or (3) in powers of c and x with possible cross terms, and our knowledge that the power series of $P(\theta)$ converges slowly, leads to the consideration of all of the following terms:

$$\text{constant}, c, x, c^2, x^2, x^3, x^4, x^5, cx, cx^2, c^2x, c^2x^2 \qquad (4)$$

The problem now is to choose from among these terms the optimum subset which contains only those terms that contribute significantly to the reduction of the residual

sum of squares. The minimum sum of squares is obtained, of course, by retaining all of the terms, but this is unsatisfactory because the t values (the estimated value of a coefficient divided by its standard deviation) become excessively small even for the terms which are known to be important. There are various approaches to this problem. A brute force method, for example, would involve selecting those terms that we "know" to be important and performing regressions over all of the remaining possible combinations of terms. In our case, we might say that the first three terms must always belong to the optimum subset and therefore calculate the regression for the 2^9 possible combinations of the remaining terms. This would be relatively costly in computer time and would still leave a large element of subjectivity in choosing the best subset(s).

Mallows[8] and Hocking and Leslie[9] have described a method for finding "good" subsets which is relatively fast and more objective by an algorithm based on the standardized total squared error, which is estimated by a statistic C_p, given by

$$C_p = \mathrm{RSS}_p/\hat{\sigma}^2 + 2p - n \tag{5}$$

where p is the number of terms in the model under consideration, $\hat{\sigma}^2$ is the residual mean square obtained by fitting the full model (which is used as an estimate of σ^2), RSS_p is the residual sum of squares associated with the p variate regression, and n is the number of data points. The "good" models will be those which have C_p about equal to p.

The Statistical Laboratory of Iowa State University has developed a program, MOUFLON, for analyzing data based on this approach. When executed in the "Hocking" mode the program will print out the several resulting subsets of terms for which $C_p \approx p$, together with various other statistical information, such as the values of the parameters and the goodness of fit.[10] This program can also analyze the data by three other useful methods, descriptively called sequential deletion, forward selection, and stepwise regression.

The first step of the program is to compute the "univariate C_p's" (the value of C_{k-1} where k is the total number of independent terms available for inclusion in the regression) for each of the possible terms. These univariate C_p's are then used to rank the terms (a large C_{k-1} means that the term is important) and then a sequential deletion procedure is used to derive the several subsets of terms that yield $C_p \approx p$. Each of these several subsets of terms gives values for the coefficients a_{kl} which yield the desired molecular parameters. The variation in the values so obtained would give a good indication of the sensitivity of the fit to inclusion or exclusion of certain terms. This variation is often larger than the standard deviations obtained by the method of linear least squares, which is due only to scatter in the data about the values predicted by a single (more or less arbitrarily selected) set of terms.

MOUFLON was used in this way to analyze the light-scattering data on seven polymer systems in order to see if any general statements could be made about the order of importance of the terms. The systems included the four samples of poly(LysZ) described below, two samples of poly(styrene) in toluene described in ref 1 and 3 (molecular weights 2.6×10^5 and 5.9×10^5), and a sample of copolyether in acetonitrile (molecular weight 9.7×10^5) for which data were supplied by Dr. W. Miller (personal communication). The information produced is extensive and need not be described here in detail. The average univariate C_p values for the seven samples, shown in Table I, are sufficient to justify a few generalizations. As might be expected, the C_p's are largest for the terms c and x, while the terms cx^2, x^4, c^2x^2, x^5, and c^2 are relatively unimportant in contributing to good fits. The constant term is always regarded as significant in this treatment. The most novel conclusion is that the cross term cx, which presumably contains some physical information which can be fairly reliably obtained, is relatively important. This was also confirmed in the observation that this term was always included in the "good" subsets of terms for individual polymer samples.

The use of MOUFLON is an approach which might yield different optimum subsets of terms for different data sets. An approach of such generality is, however, not always necessary. We have found it useful for routine analysis to have a linear least-squares program, entitled LS3, which uses the fixed set of terms

$$\text{constant, } c,\ x,\ cx,\ x^2,\ x^3, \text{ and } c^2 \tag{6}$$

The above results from MOUFLON for representative systems guided the choice of these terms, except that c^2 was retained in spite of its low univariate C_p because of other examples where it was known to be important. This Fortran program is therefore able to obtain values for all the molecular parameters $\bar{M}_w$, $\bar{\rho}_z$, A_2, A_3, R_4, R_6 (defined below), and their standard deviations. It also automatically constructs a Zimm plot showing the experimental and calculated curves by using the SIMPLOTTER subroutine developed at the Iowa State Computation Center. The plotting constant used for the abscissa of the Zimm plot is taken as $0.8/c_{max}$, where c_{max} is the maximum concentration in g/ml. As usual, the program minimizes the function

$$\mathrm{SSQ} = \sum_{i=1}^{N} w_i[y(x_i,c_i) - y_i]^2 \tag{7}$$

where y denotes Kc/R_θ, (c_i,x_i,y_i) is the ith data point of weight w_i (usually taken as one), and $y(x_i,c_i)$ is the value of y calculated from (1) using the terms in (6) with $x = x_i$ and $c = c_i$. In the linear least-squares method,[11] a set of linear equations in the a_{kl} is obtained and solved by matrix inversion. Variances and covariances of the a_{kl} are given by the elements of the inverted matrix, and are used to calculate the standard deviations of the molecular parameters using the theory of small error propagation.[12] We have used a separate program to process raw instrument readings (and to perform reflection, refraction, and volume corrections) to produce the data in a format suitable for input to LS3 and other programs. The corrections for anisotropic polarizabilities, absorption, and fluorescence are assumed to be negligible. The analysis of data by means of LS3 is described in the following sections.

Shape and Polydispersity Parameters

The coefficients of the higher powers of x in (1) contain information about the shape and polydispersity of the solute molecules. To extract this information we start with the power series expansion of $P(\theta)$ for a polydisperse system:[13]

$$P(\theta) = 1 - \frac{\beta^2}{3!}\overline{\langle r^2\rangle}x + \frac{\beta^4}{5!}\overline{\langle r^4\rangle}x^2 - \frac{\beta^6}{7!}\overline{\langle r^6\rangle}x^3 + \ldots \tag{8}$$

where $\beta = (4\pi/\lambda')$, λ' is the wavelength of light in the medium, and r is the distance between the elements of mass in a single molecule. Angular brackets denote an average over the pair distribution for a single molecule and a bar denotes an appropriate average over the molecular weight distribution. For the nth moment of the pair distribution this average is

$$\overline{\langle r^n\rangle} = \bar{M}_w^{-1}\int_0^\infty f(M)M\langle r^n\rangle \mathrm{d}M \tag{9}$$

where $f(M)$ is the weight fraction distribution of species

Table I
The Mean Univariate C_p's Obtained for Seven Representative Data Sets Listed in Order of Importance

Term	Univariate C_p	Term	Univariate C_p
c	112	cx^2	13.0
x	26	x^4	13.0
x^2	14.4	c^2x^2	12.7
cx	14.3	x^5	12.6
x^3	14.0	c^2	11.7
c^2x	13.9		

Table II
Shape Factors and Normalized Moments of the Pair Distribution Function for Various Particle Shapes

Shape	ϕ_4	ϕ_6	$\mathcal{S}$
Solid sphere	1.4286	2.4691	0.8469
Thin disk	1.6667	3.500	0.7560
Random coil	2.500	10.500	0.6720
Thin rod	2.400	7.7143	0.5580

Table III
Data Obtained from Light Scattering of Poly(LysZ) in Dimethylformamide Using LS3

Sample	$\bar{M}_w \times 10^{-5}$, g/mol	$\bar{\rho}_z$, Å	R_4	R_6	$\mathcal{S}$	$A_3 \times 10^3$, $cm^3\ g^{-2}$ mol	$A_3 \times 10^3$, $cm^6\ g^{-3}$ mol	$a_{11} \times 10^4$, $cm^3\ g^{-2}$ mol	$\left(\frac{SSQ}{n-6}\right)^{1/2} \times 10^8$	Av deviation, %
K	8.62 ± 0.39	872 ± 36	3.24 ± 0.10	21.8 ± 1.5	0.64 ± 0.08	0.19 ± 0.02	12. ± 4	0.93 ± 0.25	3.7	0.72
J	5.53 ± 0.11	591 ± 19	2.38 ± 0.30	7.0 ± 5.6	0.52 ± 0.47	0.23 ± 0.01	2.1 ± 1.4	0.84 ± 0.12	2.83	0.48
G	3.34 ± 0.19	580 ± 27	4.67 ± 0.14	44.4 ± 2.3	0.44 ± 0.05	0.24 ± 0.05	3.3 ± 5	2.2 ± 0.03	5.01	0.62
B	1.52 ± 0.02	360 ± 26	2.42 ± 1.8	−21 ± 48	Not signif	0.33 ± 0.01	4.9 ± 0.6	0.4 ± 0.2	11.2	0.67

with molecular weight M and $\bar{M}_w$ is the weight-average molecular weight. By inserting the inverse of (8) into (2) or (3) and using the coefficients defined in (1) we obtain the following molecular parameters.

$$\bar{M}_w = a_{00}^{-1} \tag{10}$$

$$\overline{\langle r^2 \rangle} = 2\bar{\rho}_z^{-2} = 6\beta^{-2}a_{01}a_{00}^{-1} \tag{11}$$

$$R_4 \equiv \overline{\langle r^4 \rangle}/\overline{\langle r^2 \rangle}^2 = \frac{10}{3}(1 - a_{00}a_{02}a_{01}^{-2}) \tag{12}$$

$$R_6 \equiv \overline{\langle r^6 \rangle}/\overline{\langle r^2 \rangle}^3 = \frac{70}{3}(1 - 2a_{00}a_{02}a_{01}^{-2} + a_{00}^2a_{03}a_{01}^{-3}) \tag{13}$$

$$A_2 = a_{10}/2 \tag{14}$$

$$A_3 = a_{20}/3 \tag{15}$$

Equations 10 and 11 express the conventional means of determining $\bar{M}_w$ and the root-mean-square radius $\bar{\rho}_z$. The dimensionless quantities R_4 and R_6 defined by (12) and (13) depend on the particle shape and polydispersity, but not on the absolute size. We consider next conditions under which one might simply separate the shape and polydispersity information in R_4 and R_6.

We assume, first, that the weight fraction is expressed by a lognormal distribution function:[14,15]

$$f(M) = \exp(-z^2/2\gamma^2)/(2\pi)^{1/2}M\gamma \tag{16}$$

where

$$z = \ln(M/\bar{M}_g) \tag{17}$$

and where γ is the polydispersity parameter and $\bar{M}_g$ is the geometric mean molecular weight. It is common practice to express the polydispersity by the ratio $\bar{M}_w/\bar{M}_n$, where n denotes the number average. It can be shown that $\bar{M}_w/\bar{M}_n = \exp(\gamma^2)$; hence $\gamma = 0$ represents a monodisperse system. The assumption of lognormality should be expected to give results well within the limits of experimental error (unless the distribution is multimodal or otherwise "ill behaved"), partly because the lognormal can be a good approximation to other unimodal, positively skewed distributions,[16] but also because distinguishing between distributions would require the measurement of the normalized third and fourth moments about the mean.[17] Even when the distribution is directly accessible, this requires rather accurate, uncensored, and untruncated data. The lognormal is adopted here primarily because it is mathematically convenient and because its properties have been well characterized.[18]

Second, we assume that all solute molecules are geometrically similar. One consequence of this assumption is that the following quantities, which are single-particle analogs of R_4 and R_6, are constants for all solute molecules:

$$\phi_4 \equiv \langle r^4 \rangle/\langle r^2 \rangle^2 \tag{18}$$

$$\phi_6 \equiv \langle r^6 \rangle/\langle r^2 \rangle^3 \tag{19}$$

Another consequence is that

$$\langle r^2 \rangle = kM^\alpha \tag{20}$$

where k and α are constants for a given shape. (For example, $\alpha = \frac{2}{3}$, 1, and 2 for solid spheres, random coils, and thin rods, respectively.) It follows that $\langle r^4 \rangle \propto M^{2\alpha}$ and $\langle r^6 \rangle \propto M^{3\alpha}$.

Defining the ith moment of the molecular weight distribution as

$$m_i \equiv \int_0^\infty M^i f(M)\mathrm{d}M \tag{21}$$

we find from (9), (12), (13), (18), and (19)

$$R_4 = \phi_4 \bar{M}_w m_{2\alpha+1}/m_{\alpha+1}{}^2 \quad (22)$$

$$R_6 = \phi_6 \bar{M}_w{}^2 m_{3\alpha+1}/m_{\alpha+1}{}^3 \quad (23)$$

For the distribution (16) we have, from (21),

$$m_i = \bar{M}_g{}^i \exp(i^2\gamma^2/2) \quad (24)$$

In particular,

$$\bar{M}_w = m_1 = \bar{M}_g \exp(\gamma^2/2) \quad (25)$$

Inserting (24) and (25) into (22) and (23) we find

$$R_4 = \phi_4 \exp(\gamma^2\alpha^2) \quad (26)$$

$$R_6 = \phi_6 \exp(3\gamma^2\alpha^2) \quad (27)$$

We define a "shape factor" $\mathbb{S}$ by eliminating the polydispersity parameter γ between (26) and (27):

$$\mathbb{S} \equiv R_6/R_4{}^3 = \phi_6/\phi_4{}^3 \quad (28)$$

Thus $\mathbb{S}$ is a dimensionless number which depends only on shape and can be determined from experimental data via (12), (13), and (28). When the shape is known, γ can be found by solving (26) or (27). This method of separating the shape and polydispersity information is applicable when the above assumptions are valid and when coefficients of terms through x^3 in (1) are accurately determined. Equation (28) holds exactly when the polydispersity is lognormal and is approximately true when it is Γ distributed. For a relatively broad distribution (e.g., $\bar{M}_w/\bar{M}_n \simeq 2$) it is not very accurate for distributions other than the lognormal, but is a good approximation for any distribution which is relatively narrow.

The values of $\mathbb{S}$ for various shapes are given in Table II, as calculated from the pair distribution function or from the series expansion of $P(\theta)$.[19] The values for $\mathbb{S}$ are seen to vary by less than a factor of 2, and therefore it is of interest to determine whether the coefficients of x^2 and x^3, required for the experimental determination of $\mathbb{S}$, can be found accurately enough in practice to be useful. Program LS3 has been designed to examine this possibility.

Application of LS3

Program LS3 has been used to analyze light-scattering data on poly(LysZ) (LysZ = ϵ-carbobenzoxy-L-lysine) in dimethylformamide solution with 4360-Å light. The experiments have been previously reported.[20] The results are given in Table III. These are somewhat at variance with the results for $\bar{M}_w$ and $\bar{\rho}_z$ obtained graphically,[20] but in most cases the discrepancies are small. The two higher molecular weight samples (K and J) give $\mathbb{S}$ values within the range given in Table I. The standard deviations should, of course, be interpreted with caution since these do not include effects of systematic errors arising either from the experiments or from the polynomial approximation, though the latter errors are small (see below).

Because of the large uncertainties often found in R_6 and $\mathbb{S}$, the method described above for obtaining shape information is not always successful. To gain added information, program LS3 includes the following analysis of the $\bar{M}_w$ and $\bar{\rho}_z$ data in terms of quantities related to the particle density for three trial shapes (rods, spheres, and coils), making use of information on polydispersity contained in R_4. We find from (9), (11), (20) and (21)

$$\bar{\rho}_z{}^2 = \frac{1}{2} k \bar{M}_w{}^{-1} m_{1+\alpha} \quad (29)$$

For thin rods $k = 1/6\delta^2$, where δ is the linear density. Thus (24), (25), and (29) give

Table IV
Comparison of $P(\theta)$ Calculated Exactly and by Third Degree Polynomial Approximation for Random Coils of Various Sizes

	$s = 1$		$s = 2$		$s = 3$	
θ, deg	Exact	Approx	Exact	Approx	Exact	Approx
30	0.9780	0.9780	0.9164	0.9163	0.8260	0.8257
50	0.9430	0.9430	0.7990	0.7983	0.6253	0.6103
70	0.8988	0.8988	0.6747	0.6677	0.4592	0.3110
90	0.8522	0.8521	0.5677	0.5333	0.3468	−0.331
110	0.8093	0.8088	0.4865	0.3834	0.2765	−1.64
130	0.7743	0.7732	0.4304	0.2132	0.2340	−0.3966

$$\delta = \frac{\bar{M}_w}{12^{1/2}\bar{\rho}_z} \exp(3\gamma^2/2) \quad (30)$$

For solid spheres $k = (6/5)(3/4\pi d)^{2/3}$, where d is the density. Thus

$$d = \left(\frac{3}{5}\right)^{3/2} \left(\frac{3}{4\pi}\right) \frac{\bar{M}_w}{\bar{\rho}_z{}^3} \exp(5\gamma^2/6) \quad (31)$$

For random coils $k = \alpha_e{}^2 b^2/3M_0$, where α_e is the expansion coefficient, b is the effective length per segment, and M_0 is the molecular weight per segment.[21] Thus

$$\alpha_e b = (6M_0/\bar{M}_w)^{1/2} \bar{\rho}_z \exp(-\gamma^2/2) \quad (32)$$

The value of γ is calculated for each of the shapes from the measured value of R_4 using (26). As an illustration, the data for poly(LysZ) sample K give δ(rod) = 319 dalton/Å, d(sphere) = 1.12×10^{-3} g/cm^3, and $\alpha_e b$(coil) = 28.8 Å, using $M_0 = 263$. The value of δ is of a reasonable magnitude, while those for d and $\alpha_e b$ are unrealistic, suggesting that the molecules more nearly resemble a rod than a sphere or random coil. It was found previously[20] that a flexible rod shape best fit the light scattering and other data for poly(LysZ). The present results are not inconsistent with this conclusion. The value of $\mathbb{S}$ found for sample K is intermediate between the rod and coil values (Table II), though the experimental uncertainty is too large to permit an estimate of the degree of flexibility from this quantity alone.

In order to explore the effects of noise, polydispersity, and size upon the accuracy of parameters obtained by LS3, a program was also written to generate *simulated* light-scattering data based on the exact expression for $P(\theta)$ rather than the first few terms of its series expansion. The program carries out the integration in the calculation of $P(\theta)$ numerically for the shapes rod, coil, or sphere using the lognormal distribution for the molecular weight and superimposes any desired level of Gaussian noise. Data generated by this program using various molecular parameters, shapes, and noise levels were then analyzed by LS3 and the molecular parameters given as output were compared to those used in the simulation. As expected, the parameters A_2, A_3, and a_{11} were always given accurately when the noise was negligible, but systematic errors appeared in the coefficients of the powers of x^n ($n = 0$ to 3) when the size of the particles increased beyond a certain point. The reason for this is simply that higher terms that have been omitted in eq 8 become important for large particles. The power series is accurate only when the particle size and/or the angle θ is sufficiently small. This effect has been clearly described previously.[4] On the other hand, the various forms for $P(\theta)$ can be distinguished only for data in which terms at least through x^2 are significant.

Since terms only through x^3 are considered in LS3, the effect of the slow convergence of $P(\theta)$ on the values obtained for $\bar{M}_w$ and $\bar{\rho}_z$ is worthy of consideration. Table IV

Table V
Results Obtained by Fitting Simulated Data for Monodisperse Random Coils by Program LS3

s (true)	s (found)	$\bar{M}_w$	$\bar{\rho}_z$, Å	R_4	R_6	S
1.0	0.9997	199998	260.0	2.4865	10.05	0.6537
3.0	3.080	201665	801.0	2.770	14.87	0.7000
5.0	6.010	226032	1563	3.249	22.12	0.6451
7.0	9.696	270187	2522	3.3221	23.17	0.6321
9.0	13.596	310371	3536	3.3313	23.30	0.6304
11.0	17.434	337752	4534	3.3328	23.33	0.6301
True values:		200000	260s (true)	2.500	10.500	0.6720

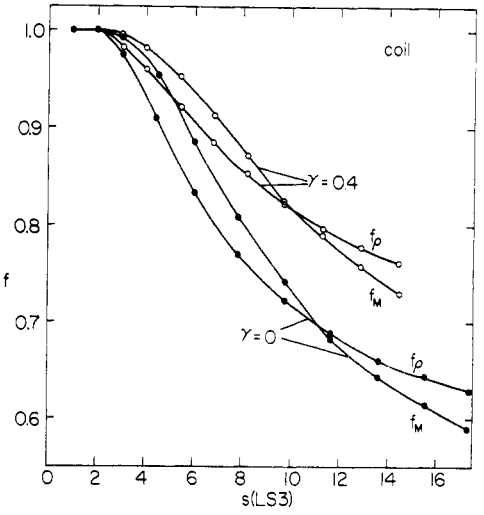

Figure 1. Correction factors for $\bar{M}_w$ and $\bar{\rho}_z$ vs. size parameter for coils with polydispersity $\gamma = 0$ and 0.4, applicable to measurements in the 30–150° angular range. The size parameter s given by LS3 is used as the abscissa while the true values used in the simulation were integers from 1 to 11 at points indicated by circles. No noise was added to the simulated data.

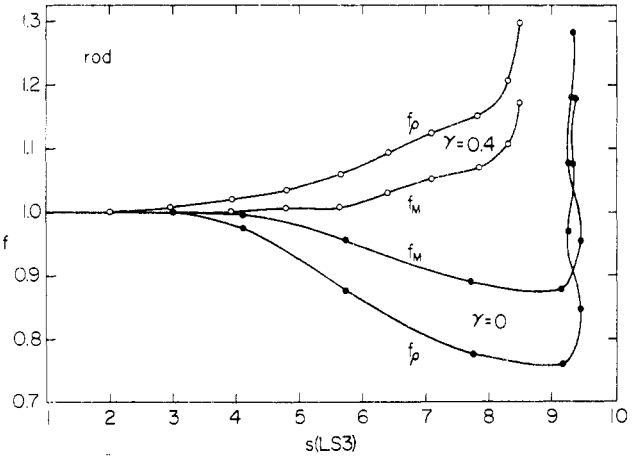

Figure 2. Same as Figure 1 for thin rods.

shows a comparison of values of $P(\theta)$ for a random coil calculated by the exact formula and by the expansion in eq 8 with terms through x^3. The size is expressed by a dimensionless size parameter s defined by

$$s = 4\pi g(\rho/\lambda') \tag{33}$$

where ρ is the single-particle root-mean-square radius and g has the value $\frac{5}{3}^{1/2}$ for solid spheres, $2^{1/2}$ for thin disks, 1 for random coils, and $3^{1/2}$ for thin rods. Exact expressions for $P(\theta)$ expressed as functions of $sx^{1/2}$ are available for these shapes.[13] From the results in Table IV we would expect LS3 to give a good fit with accurate values of the molecular parameters when $s \leq 1$, while for $s > 2$ systematic errors may be introduced due to the neglect of terms higher than x^3.

The extent of this error in the various parameters is seen in Table V, which gives the results obtained from LS3 applied to simulated light-scattering data on monodisperse random coils of various sizes. The data were calculated using the exact form of $P(\theta)$ at intervals of 20° in the range $30° \leq \theta \leq 150°$. As expected, the errors become significant for s greater than about 2. If one had data for particles of unknown size and polydispersity but known shape one could conceivably use results such as these to correct the values found by LS3 in a convenient way through a "correction factor", f, defined by

$$f = (\text{true value of parameter})/(\text{value found by LS3}) \tag{34}$$

This quantity is plotted in Figure 1 for coils and in Figure 2 for rods as a function of the value of s found by LS3. If the polydispersity can be estimated, then use of the correction factors can result in considerable improvement in the estimates of the molecular parameters obtained by LS3. The correction factors in Figures 1 and 2 apply strictly only to data which cover the angular range 30–150°. This is true regardless of the wavelength of light used because the particle scattering factor is a function of only $s \sin(\theta/2)$. The systematic error introduced by LS3 results from the slow convergence of the power series for $P(\theta)$ and from the absence of data at angles sufficiently low to establish the lower terms in the series more accurately. If data at lower angles were available, the correction would be correspondingly smaller. These results also demonstrate the well-known fact that the use of longer wavelengths for large particles can reduce systematic errors in the quantities which depend only on the lower terms in the expansion of Kc/R_θ. The values given by LS3 for the shape factor S for these same simulated data sets are shown in Figure 3.

The use of correction factors as described here is limited to cases where the shape is known so that a comparison with simulated data can be made. Perhaps a more generally useful conclusion from this study of simulated data is a statement of the limits on particle size below which a program such as LS3 produces a systematic error less than a desired limit. Table VI shows these limits for errors in $\bar{M}_w$ and $\bar{\rho}_z$ for rods and coils when scattering measurements cover the angular range 30–150°. It is seen that $\bar{\rho}_z/\lambda'$ must be less than a few tenths to obtain reasonable accuracy in $\bar{M}_w$ and $\bar{\rho}_z$, and that the size limits are larger for polydisperse systems than for monodisperse systems. For the poly(LysZ) samples in Table III $\bar{\rho}_z/\lambda'$ does not exceed 0.28. Thus the systematic error in $\bar{M}_w$ and $\bar{\rho}_z$ for these samples is less than 5% for either a rod or a coil shape.

Table VI
Values of $\bar{\rho}_z/\lambda'$ below which the Errors in $\bar{M}_w$ and $\bar{\rho}_z$ Obtained by LS3 are Less Than the Specified Limits

Error limit, %	Coil				Rod			
	$\gamma = 0$		$\gamma = 0.4$		$\gamma = 0$		$\gamma = 0.4$	
	M	ρ	$\bar{M}_w$	$\bar{\rho}_z$	M	ρ	$\bar{M}_w$	$\bar{\rho}_z$
5	0.32	0.25	0.41	0.32	0.24	0.21	0.36	0.26
10	0.39	0.32	0.51	0.43	0.26	0.23	0.37	0.33
20	0.49	0.45	0.70	0.71		0.27	0.47	0.45

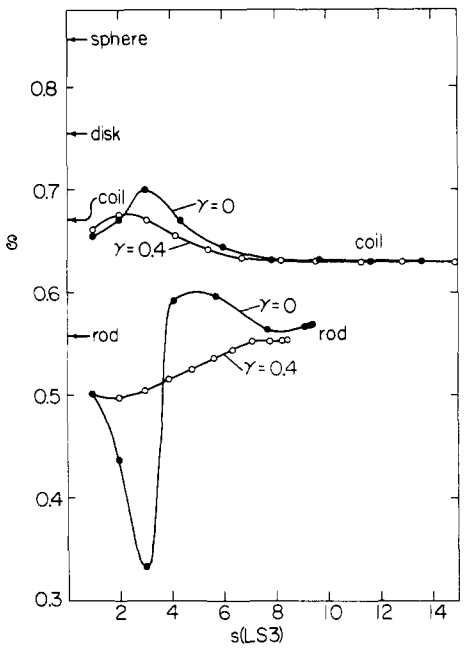

Figure 3. Shape factor given by LS3 vs. size parameter given by LS3 for coils and rods. The same simulated data as in Figures 1 and 2 were used. The true values of S are shown on the ordinate for the common shapes.

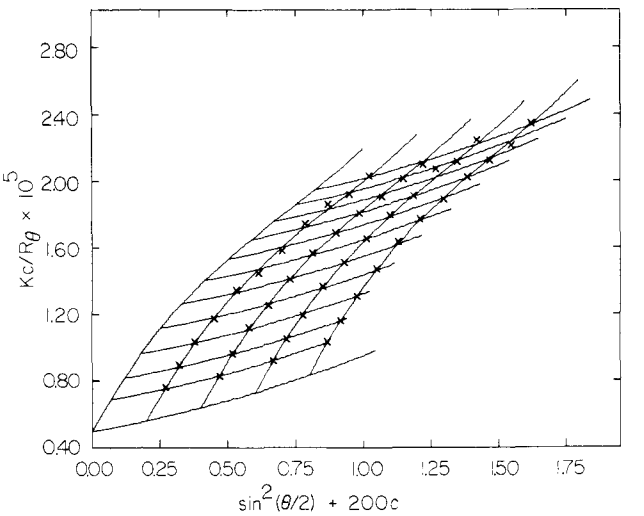

Figure 4. Zimm plot generated by LS3 for simulated data for polydisperse rods with $s = 8$, $\gamma = 0.4$, random noise of 0.5% in Kc/R_θ. Assumed coefficients: $a_{10} = 4.00 \times 10^{-4}$ cm^3 g^{-2} mol; $a_{20} = 0.100$ cm^6 g^{-3} mol; $a_{11} = 3.00 \times 10^{-4}$ cm^3 g^{-2} mol.

Figure 4 shows a Zimm plot generated by LS3 applied to simulated data for polydisperse rods with $s = 8$ and $\gamma = 0.4$. An rms random noise level of 0.5% was introduced in the ordinate. The smooth curves were calculated using the least-squares parameters. Even though a good fit to the data is obtained, it is seen from Figure 2 that there is a systematic error of 8% in $\bar{M}_w$ and 16% in $\bar{\rho}_z$ for this system. The standard deviations found by LS3 are 4.0% for $\bar{M}_w$ and 3.6% for $\bar{\rho}_z$.

Nonlinear Least Squares

A method of avoiding the whole problem of the slow convergence of the power series expansion of $P(\theta)$ for large θ or large particle size is to use a nonlinear least-squares model

Table VII
Parameters Obtained by NLN Applied to Simulated Light-Scattering Data for Polydisperse Rods

	Trial shapes														
	Rod					Coil					Sphere				
	$10^{-6}a_{00}$	s	γ	10^{12}SSQ	Av dev, %	$10^{-6}a_{00}$	s	γ	10^{12}SSQ	Av dev, %	$10^{-6}a_{00}$	s	γ	10^{12}SSQ	Av dev, %
I															
Input	4.84	8.22	0.37	3.9		5.20	4.14	0.64	1265		5.40	3.60	2.0	36	
Output	4.95	8.01	0.39	0.3	0.38	6.13	3.89	0.97	4.7	2.1	5.43	4.18	2.5	17.1	4.0
True values	5.00	8.00	0.40	Noise = 0.5%											
II															
Input	4.50	8.40	0.32	2.5		5.40	4.00	0.90	14.4		5.40	3.30	2.0	132	
Output	5.09	7.94	0.38	0.3	0.36	6.17	3.66	0.93	4.8	2.1	5.41	4.15	2.5	17.3	4.0
True values	5.00	8.00	0.40	Noise = 0.5%											
III															
Input	4.71	8.36	0.35	7.5		4.99	4.27	0.64	1182		4.96	3.31	1.48	590	
Output	4.91	7.55	0.32	1.1	0.73	5.88	3.92	0.94	4.8	2	5.36	4.24	2.47	16.6	3.9
True values	5.00	8.00	0.40	Noise = 1%											
IV															
Input	5.09	5.66	0.35	0.29		5.09	3.27	0.66	232		5.09	2.53	1.50	405	
Output	5.05	5.84	0.38	0.002	0.05	5.46	3.24	0.94	0.8	1.0	6.18	2.78	1.55	7.4	3.1
True values	5.00	6.00	0.40	Noise = 0											
V															
Input	5.50	4.40	0.44	66		5.02	2.26	0.69	20		5.02	1.75	1.53	149	
Output	5.04	3.91	0.38	0.002	0.05	5.06	2.19	0.87	0.1	0.1	5.13	2.38	1.62	0.95	1.3
True values	5.00	4.00	0.40	Noise = 0											

fitting approach which would use the various exact forms of $P(\theta)^{-1}$ to calculate and minimize the sum of squares SSQ of the difference between the experimental data and the calculated values. We have written such a Fortran program, called NLN, which is able to carry out the least-squares fits for three representative shapes (spheres, rods, and coils) which are lognormally disperse. Because of the complicated form of $P(\theta)^{-1}$, which involves an integration over the polydispersity, it is not possible to utilize the gradient of the function being minimized in order to find the minimum value and the associated molecular parameters. A slower algorithm is unavoidable and we use the subroutine ZXPOWL,[22,23] which finds a local minimum of a function of N variables.

Since LS3 finds the coefficients of c, c^2, and cx with sufficient accuracy it is necessary only to vary $\bar{M}_w$, $\bar{\rho}_z$, and γ to carry out the minimization. The program NLN reads the experimental data, trial values of the parameters $1/\bar{M}_w$, s, and γ for each of the shapes to be tested, and a_{10}, a_{20}, and a_{11} obtained by LS3. The number of iterations to be carried out and a number used to stop the iterations if the improvement in the fit is not increasing significantly must also be specified as input. At each iteration the current best value of the parameters and SSQ is printed out so that the convergence can be followed. About 10 iterations generally suffice to decrease SSQ and change the parameters to values sufficiently close to the best values that little is gained by further iterations. The cost is several times as large as for LS3. Although one might expect to distinguish the various shapes for the larger values of s and to derive estimates of the polydispersity by a program of this sort, it should be recognized that it has all of the usual disadvantages of any nonlinear program: it is more costly, requires reasonably good estimates of the input parameters, and can find false minima which may require some judgement to recognize.

Table VII shows some results of the use of NLN to "deduce" the shape and polydispersity of the scattering particles using simulated experimental data with various noise levels for polydisperse rods of various sizes. Examples I and II were carried out on the data shown in Figure 4, using different sets of input trial values. As expected, the fit is much better for the rod than for the coil or sphere, judging from the magnitudes of SSQ or the average deviations obtained with the output parameters. These examples show that the trial parameters converge to values very close to the correct values when the trial shape is correct, and that this convergence is relatively insensitive to the input parameters. Example III shows that it is still possible to distinguish the shape on the basis of goodness of fit when the noise level is doubled to 1% in the same system treated in examples I and II. Examples IV and V illustrate results for smaller particles (smaller s) with no noise in the simulated data. The goodness of fit is best for the rod shape in these cases as well; however, the SSQ and average deviations for the coil and sphere trial shapes would not be considered large for typical experimental data, and it seems likely that noise would tend to obliterate the distinction among shapes by this approach in these cases.

Acknowledgment. This investigation was supported by a research grant from the National Institute of General Medical Sciences (GM-13684).

Supplementary Material Available. Listings of Fortran programs LS3 and NLN will appear immediately following this article in the microfilm edition of this volume of the journal. Photocopies of the supplementary material from this paper only or microfiche (105 × 148 mm, 24× reduction, negatives) containing all of the supplementary material for the papers in this issue may be obtained from the Journals Department, American Chemical Society, 1155 Sixteenth St., N.W., Washington, D.C. 20036. Remit check or money order for $4.50 for photocopy or $2.50 for microfiche, referring to code number MACRO-75-858.

References and Notes

(1) J. M. Evans, M. B. Huglin, and J. Findley, *J. Appl. Polym. Sci.*, **2**, 2159 (1967).
(2) W. A. J. Bryce, *Polymer*, **10**, 803 (1969).
(3) W. Miller and R. F. T. Stepto, *Eur. Polym. J.*, **7**, 65 (1971).
(4) P. F. Mijnlieff and D. S. Coumou, *J. Colloid Interface Sci.*, **27**, 553 (1967).
(5) J. M. Evans, "Light Scattering from Polymer Solutions", M. B. Huglin, Ed., Academic Press, New York, N.Y., 1972, p 89.
(6) B. H. Zimm, *J. Chem. Phys.*, **16**, 1093 (1948).
(7) P. J. Flory and A. M. Bueche, *J. Polym. Sci.*, **27**, 219 (1958).
(8) C. Mallows, "Choosing a Subset Regression", paper presented at the Joint Statistical Meeting, Los Angeles, Calif., 1966.
(9) R. R. Hocking and R. N. Leslie, *Technometrics*, **9**, 531 (1958).
(10) B. L. Hanson, "MOUFLON Reference Manual", Statistical Laboratory, Iowa State University, Ames, Iowa, 1969.
(11) P. R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences", McGraw-Hill, New York, N.Y., 1969, Chapter 9.
(12) Reference 11, Chapter 4.
(13) E. P. Geiduschek and A. Holtzer, *Adv. Biol. Med. Phys.*, **6**, 431 (1958).
(14) H. Wesslau, *Makromol. Chem.*, **20**, 111 (1956).
(15) R. Chiang, *J. Polym. Sci.*, **36**, 99 (1959).
(16) A. L. Koch, *J. Theor. Biol.*, **23**, 251 (1968).
(17) K. Pearson, *Phil. Trans. R. Soc. London*, **A216**, 456 (1916).
(18) P. Yuan, *Ann. Math. Stat.*, **4**, 30 (1933).
(19) P. Kratochvil, ref 5, p 333.
(20) J. Applequist and P. Doty, "Polyamino Acids, Polypeptides, and Proteins", M. A. Stahmann, Ed., University of Wisconsin Press, Madison, Wis., 1962, p 161.
(21) C. Tanford, "Physical Chemistry of Macromolecules", Wiley, New York, N.Y., 1961, p 168.
(22) International Mathematical Sciences Libraries, Inc., "Library 1 Reference Manual", 4th ed, Houston, Texas, 1973.
(23) W. Zangwill, *Comput. J.*, **10**, 293 (1967).