Tetrahedron Vol 37, pp 2115 to 2119, 1981
Printed in Great Britain

0040-4020/81/112115-05$02.00/0

RELIABILITY OF ASSOCIATION CONSTANTS OF 1:1 MOLECULAR COMPLEXES FROM SPECTROPHOTOMETRIC DATA

G. CARTA, G. CRISPONI and V. NURCHI
Istituto Chimico Policattedra, Università di Cagliari, Via Ospedale 72, 09100 Cagliari, Italy

(*Received U.K. 4 August* 1980)

Abstract—Some basic factors affecting the reliability of K (association constant) and ϵ (molecular extinction coefficient) estimates for 1:1 molecular complexes are studied. To accomplish this, the error matrix is examined and a new variable $G = K^{-1}((a+b+K^{-1})^2 - 4ab)^{-1/2}$ (where a and b are the concentrations of the reagents) is introduced. Whenever G has the same value for all the experimental points, K and ϵ are undetermined. A great dispersion of G values, on the contrary, contributes to more reliable estimates. The variable G has in fact the same role as the independent variable in a linear relationship; therefore the problem of the optimal choice of experimental points is reduced to one of a simple linear regression.

Since the Benesi and Hildebrand[1] work on 1:1 molecular complexes formation constants, several authors have displayed a great interest in this problem[2-5] on account of the simplicity and accuracy of spectrophotometric methods, as well as the facility of the approximate graphical and/or linear methods proposed. The reliability of association constants and molar extinction coefficients calculated with such methods has been critically discussed by Person,[6] who emphasizes that more accurate values are obtained when the complex equilibrium concentration is of the same order of magnitude as the equilibrium concentration of the more dilute component. Derenleau[7] introduced the saturation fraction concept, and pointed out that 75% of the saturation curve is required to verify a theoretical model.

Subsequently, La Budde and Tamres,[8] by assuming only a 1:1 complex formation, remarked how the error limits in a linear regression depend both on the number of data points and on the experimental errors, and also particularly on the range covered by the independent variable. More recently, Christian *et al.*[9] affirmed that good estimates of K and ϵ parameters can be obtained by properly weighing the observations in the linear regression.

With the increasing availability of digital computers, several interesting publications based on non-linear least squares methods[10-16] have recently appeared. None of these authors, unlike Person, Derenleau and La Budde *et al.* pays special attention to the concentration range.

Discordant K and ϵ estimates have been reported.[17] In order to clarify these anomalies, deviations from the simple 1:1 association have been invoked, such as:

(i) formation of higher order complexes;[18]
(ii) variation with the concentration of the reagents of the molar extinction coefficient of the complex;[19]
(iii) interaction between reagents and solvent.[20]

The aim of our work is to contribute to the analysis of the factors, other than (i), (ii) and (iii) affecting the K and ϵ parameters, estimated by a general least squares method. In the following discussion therefore it has been assumed that:

(i) only a 1:1 molecular complex is formed by interaction between the two reagents

$$A + B \overset{K}{\rightleftharpoons} X,$$

(ii) the absorption is due to the complex alone, and Beer's law is rigorously valid at some fixed wavelength,
(iii) the relative error of absorbance $\Delta D_e/D_e$ is drawn from normally distributed populations with $\mu = 0$ and the same variance σ^2 in the whole complex concentration range.

It will be rigorously shown how a non-proper choice of the concentration range of the reagents can lead to non-significant K and ϵ estimates, even if there is no deviation from the assumed simple model.

General equations

The K and ϵ values with their respective standard deviations can be obtained by the least squares method, by which we can find the couple (K_M, ϵ_M) that minimizes the expression:

$$X^2 = \sum^{N} (D_e - D)^2 \cdot W \qquad (1)$$

where the sum is over all the N experimental points $(D_{se}, a_i, b_i, i = 1, N)$. D_e and D are respectively the experimental and calculated absorbances, for unit path-length. D is given by the following relations:

$$D = \epsilon x \qquad (2)$$

$$x = 0.5((a+b+K^{-1}) - ((a+b+K^{-1})^2 - 4ab)^{1/2}) \qquad (3)$$

where the sum is over all the N experimental points (D_e, two reagents A and B, x is the complex concentration calculated by (3) for a given K value, and W is the weight of the single measure.

K and ϵ are completely indeterminate whenever the concentrations of the reagents are such that the relative

variations of x values with K are the same for all the N solutions; then, in fact, the same set of N values of D can be obtained for whatever K, compensating the variations in x values with the right ϵ value. In such a situation the two parameters are completely correlated and an infinite number of K and ϵ couples give the same minimum value of the expression (1). The above condition is achieved when all N points have the same value of a variable G defined as:

$$G = \lim (\Delta x/x)/(\Delta K/K) = \frac{\partial x}{\partial K}\frac{K}{x}$$
$$\Delta K \to 0 = K^{-1}((a+b+K^{-1})^2 - 4ab)^{-1/2}. \quad (4)$$

In Fig. 1, we plot the contour lines of G, which can only assume values between 0 and 1.

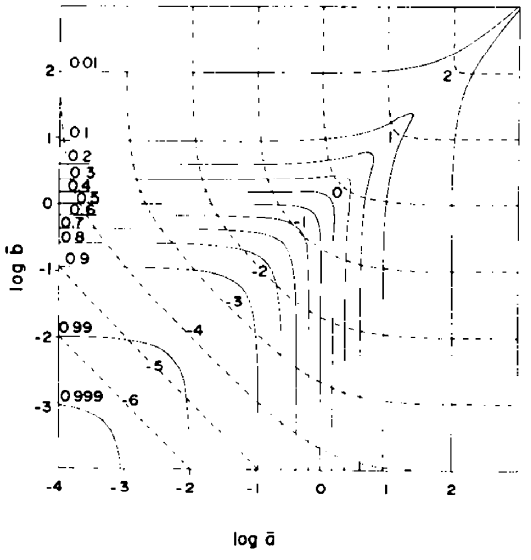

Fig. 1. The contour lines for different G's are plotted as full lines in the (log $\bar{b}$, log $\bar{a}$) plane. Dotted lines represent points having the same concentration of the complex; the numbers on these lines are log $\bar{x}$; $\bar{a}$, $\bar{b}$ and $\bar{x}$ are the concentrations in 1/K units (i.e. $\bar{a} = aK$, $\bar{b} = bK$ and $\bar{x} = xK$) in order to make the figure independent of the K values. For a data set the more the G values are scattered, the more the determination of K and ϵ is precise. On the contrary when all the points lie on the same contour line the estimates of K and ϵ are completely indeterminate.

In order to evaluate K and ϵ and their reliability, the Gauss–Newton linearization method has been employed.

The absorbances D in relation (1) can be expressed as a function of the parameter increments $\Delta\epsilon_0 = \epsilon - \epsilon_0$ and $\Delta K_0 = K - K_0$ (where K_0 and ϵ_0 are the values at starting point), by a Taylor series truncated to the first term:

$$X^2 = \sum^N (D_e - D)^2 W = \sum^N (D_e - D_0 - \Delta K_0 \cdot D'_{K_0} - \Delta\epsilon_0 \cdot D'_{\epsilon_0})^2 W \quad (5)$$

where $D'_{K_0} = (\partial D/\partial K)_0 = (\epsilon_0/K_0)x_0 G_0$, and $D'_{\epsilon_0} = (\partial D/\partial \epsilon)_0 = x_0$ and subscript 0 indicates values calculated with $K = K_0$ and $\epsilon = \epsilon_0$.

The eqn (5) is minimized with respect to ΔK_0 and $\Delta\epsilon_0$, equating to zero the derivatives

$$\frac{\partial X^2}{\partial \Delta K_0} = 0 \qquad \frac{\partial X^2}{\partial \Delta \epsilon_0} = 0$$

The resulting equation system can be solved by matrix inversion; K and ϵ values of increasing precision, with the relative variance-covariance matrix, should be obtained by an iterative process.

The diagonal elements C_{kk} and $C_{\epsilon\epsilon}$ of inverse matrix allow a calculation of the variances of the parameter estimates K and ϵ according to

$$S_k^2 = S^2 \cdot C_{kk} \quad (7a)$$

$$S_\epsilon^2 = S^2 \cdot C_{\epsilon\epsilon} \quad (7b)$$

where $S^2 = X^2/(N-2)$, and $N-2$ are the degrees of freedom for the estimate of S^2.

From the off-diagonal element $C_{k\epsilon}$, we can obtain the correlation coefficient $r_{k\epsilon} = C_{k\epsilon}(C_{kk}C_{\epsilon\epsilon})^{-1/2}$, i.e. the degree to which possible errors on one parameter affect the other.

The analytical expression of the inverse matrix elements is given by:

$$C_{kk} = \sum^N (D'_\epsilon)^2 \cdot W/\mathrm{Det} = \left(\frac{K}{\epsilon}\right)^2 \frac{\sum^N x^2 W}{\sum^N x^2 W \sum^N x^2 G^2 W - (\sum^N x^2 G W)^2} \quad (9a)$$

$$C_{\epsilon\epsilon} = \sum^N (D'_k)^2 \cdot W/\mathrm{Det} = \frac{\sum^N x^2 G^2 W}{\sum^N x^2 W \sum^N x^2 G^2 W - (\sum^N x^2 G W)^2} \quad (9b)$$

$$C_{k\epsilon} = \sum^N D'_k \cdot D'_\epsilon \cdot W/\mathrm{Det} = \frac{K}{\epsilon}\frac{\sum^N x^2 G W}{\sum^N x^2 W \sum^N x^2 G^2 W - (\sum^N x^2 G W)^2}. \quad (9c)$$

Equations (9) are calculated with those K and ϵ values for which X^2 is minimum, and $\mathrm{Det} = \sum^N (D'_k)^2 \cdot W \sum^N (D'_\epsilon)^2 \cdot W - (\sum^N D'_k \cdot D'_\epsilon)^2 \cdot W^2$.

If $W = 1/D_e^2$ (which implies that the relative errors of absorbance are minimized, and that condition (iii) at pag. 3) is satisfied) and $D_e = \epsilon x$ (which neglects experimental errors) are used the eqns (9) become

$$C_{kk} = K^2/\sum^N (G - \bar{G})^2 \quad (10a)$$

$$C_{\epsilon\epsilon} = \epsilon^2 \overline{G2}/\sum^N (G - \bar{G})^2 \quad (10b)$$

$$C_{k\epsilon} = \epsilon K \bar{G}/\sum^N (G - \bar{G})^2 \quad (10c)$$

where $\bar{G} = \sum^N G/N$ and $\overline{G2} = \sum^N G^2/N$

The eqns (10), divided by K^2, ϵ^2 and $K\epsilon$ respectively, will look like the analytical expressions of inverse matrix elements in a linear regression,[23] where G is the independent variable.

Calculations and results

Three sets of simulated data, denoted as *a*, *b* and *c* respectively, were examined to check equations (10). Each set consists of 12 points with concentrations of reagents (concentration errors were considered negligible) chosen is such a way that half of the points assume a particular value of G, and the other half another. This gathering at each end of the G interval is chosen because it is that which maximizes $\sum^{N}(G-\bar{G})^2$ in (10), and so the result for the three sets can be compared, depending only on the G range.

The complex absorbances D_e have been calculated for $K = 1$ (moles/l)$^{-1}$, $\epsilon = 100$ (moles/l)$^{-1}$ cm^{-1} and for a unit path length cell; then errors were added to these values in such a way that their relative errors were normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 0.01^2$, by a program based on random number generation.

Table 1 reports the concentrations of the reagents and simulated absorbances for the three data sets, while Table 2 shows the results, obtained with a program based on Gauss–Newton linearization method, using $W = 1/D_e^2$. In Table 3 we report the values of C_{kk}, $C_{\epsilon\epsilon}$, $r^2_{k\epsilon}$, calculated using (10), for 45 sets of 12 data, each set with a particular G range and distribution similar to those of *a*, *b*, *c*.

†These conditions imply that the model is known; if not so the Derenleau's considerations (i.e. that 75% of the saturation curve is required to verify a theoretical model) are valid.

Both C_{kk} and $C_{\epsilon\epsilon}$ decrease as $\sum^{N}(G-\bar{G})^2$ (i.e. as the G's dispersion) increases, while $C_{\epsilon\epsilon}$ depends also on $\overline{G2}$; in every case we have a better relative precision for ϵ than for K, because $\overline{G2}$ is always smaller than 1. As for as $r^2_{k\epsilon}$ is concerned generally we can say that, $\sum^{N}(G-\bar{G})^2$ being equal, lower values of $r^2_{k\epsilon}$ mean a better selection of experimental points.

The same conclusions can be reached in another way, which fact can be shown graphically in Figs. 2 and 3, where we report the minimum values of X^2 as a function of K and ϵ respectively.

DISCUSSION

The relations (7) show that the K and ϵ variances depend on two factors: experimental errors S^2 and C_{ii}'s, whose analytical expressions are given in (9) and (10). S^2 being fixed on the basis of experimental conditions, we are able to affect only C_{ii}'s, minimizing them in the limits of the experimental restraints. If the conditions are such that the model implied in (10) is valid, this means maximizing $\sum^{N}(G-\bar{G})^2$ by making the G range as great as possible and by selecting the points at the outer limits.†

The confidence limits for the parameters can be calculated from standard deviations using the proper t multipliers

$$K \pm t_\alpha \cdot S_k \qquad (11a)$$

$$\epsilon \pm t_\alpha \cdot S_\epsilon \qquad (11b)$$

where t_α has the same number of degrees of freedom as S, and a confidence coefficient $(1-2\alpha)$. A $(1-\alpha)$ confidence region for the two simultaneous estimates of

Table 1. The data for the three simulated cases *a*, *b* and *c* are shown; a and b are the concentrations of the reagents, D_e is the simulated absorbance and the corresponding G values are indicated in brackets. Each set has only two extreme G values; this particular distribution is not necessary but it makes the treatment easier

CASE *a*				CASE *b*				CASE *c*			
a x 10^3	b x 10^3		D_e	a x 10^3	b x 10^3		D_e	a x 10^3	b x 10^3		D_e
1.995	9002.	(.1)	.180	1.995	9002.	(.1)	.179	7.944	245.2	(.8)	.159
3.162	9003.	"	.285	3.162	9003.	"	.280	12.59	242.3	"	.240
6.310	9005.	"	.572	6.310	9005.	"	.565	19.95	237.6	"	.376
12.59	9010.	"	1.12	12.59	9010.	"	1.14	50.12	217.2	"	.850
25.12	9020.	"	2.28	25.12	9020.	"	2.24	100.0	178.2	"	1.41
50.01	9040.	"	4.49	50.01	9040.	"	4.45	158.5	122.1	"	1.51
12.59	100.8	(.9)	.115	2.512	4002.	(.2)	.199	12.59	100.8	(.9)	.114
15.95	98.05	"	.139	3.981	4002.	"	.314	15.85	98.06	"	.140
19.95	94.55	"	.169	7.944	4005.	"	.636	19.95	94.55	"	.168
39.81	76.79	"	.275	15.85	4009.	"	1.27	39.81	76.79	"	.279
50.12	67.02	"	.295	31.63	4019.	"	2.57	50.12	67.02	"	.289
63.10	54.14	"	.303	63.10	4038.	"	5.06	63.10	54.14	"	.302

Table 2. The estimates of the two parameters K_M and ϵ_M are reported, with the values S, S_k/K and S_ϵ/ϵ. In brackets the corresponding values, obtained using te eqn (10) for K = 1 and ϵ = 100, are reported

K	ϵ	S	S_k/K	S_ϵ/ϵ
.992	100.2	.009 (.010)	.006 (.007)	.004 (.004)
1.06	98.8	.009 (.010)	.056 (.059)	.008 (.009)
1.03	97.6	.013 (.013)	.073 (.076)	.062 (.064)

Table 3. Values of inverse matrix elements C_{kk}, $C_{\epsilon\epsilon}$ and $r^2_{k\epsilon}$ ($r_{k\epsilon}$ is the correlation coefficient) for different G ranges, calculated by (10) using K = 1 and ϵ = 100. It is clear the C_{kk} depends only on the G range, while $C_{\epsilon\epsilon}$ depends also on the actual G values; $C_{\epsilon\epsilon}$ is always smaller than the corresponding C_{kk} thus showing a better relative precision in ϵ. The dashes indicate an infinite value

G/G	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	----				C_{kk}				
.2	33.3	----							
.3	8.33	33.3	----						
.4	3.70	8.33	33.3	----					
.5	2.08	3.70	8.33	33.3	----				
.6	1.33	2.08	3.70	8.33	33.3	----			
.7	.926	1.33	2.08	3.70	8.33	33.3	----		
.8	.680	.926	1.33	2.08	3.70	8.33	33.3	----	
.9	.521	.680	.926	1.33	2.08	3.70	8.33	33.3	----
.1	----				$C_{\epsilon\epsilon}$≠				
.2	.833	----							
.3	.417	2.17	----						
.4	.315	.833	4.17	----					
.5	.271	.537	1.41	6.83	----				
.6	.247	.417	.833	2.17	10.2	----			
.7	.232	.353	.604	1.20	3.08	14.2	----		
.8	.221	.315	.487	.833	1.65	4.17	18.8	----	
.9	.214	.289	.417	.647	1.10	2.17	5.42	24.2	----
.1	1.00				$\gamma^2_{k\epsilon}$				
.2	.900	1.00							
.3	.800	.961	1.00						
.4	.735	.900	.980	1.00					
.5	.693	.844	.939	.988	1.00				
.6	.662	.800	.900	.962	.992	1.00			
.7	.640	.764	.863	.930	.973	.994	1.00		
.8	.626	.736	.829	.900	.949	.980	.996	1.00	
.9	.610	.711	.800	.871	.924	.962	.984	.996	1.00

≠ These values are made independent of the ϵ value by dividing $C_{\epsilon\epsilon}$ by ϵ^2 (ϵ = 100).

K and ϵ (with true values K_0 and ϵ_0) can be evaluated by the following relation:

$$C_{\epsilon\epsilon}(K-K_0)^2 - 2C_{k\epsilon}(K-K_0)(\epsilon-\epsilon_0) + C_{kk}(\epsilon-\epsilon_0)^2 = 2FS^2(C_{kk}C_{\epsilon\epsilon} - C_{k\epsilon}^2) \quad (12)$$

where F is the F-distribution with 2 and (N − 2) degrees of freedom and a significance level of α. This equation defines ellipses centered at the point K_0, ϵ_0 with inclined axes, and is rigorously valid for linear models; it can also be applied to non-linear systems to the extent to which the linearized form is a good approximation of the true model.[22] The ellipses for the cases *a*, *b* and *c* are shown in Fig. 4. The areas of the ellipses are always smaller than those of the rectangles whose sides are given by (11), also for the same significance level. In fact, to regard the confidence intervals given by (11)† separately is incorrect because of the high correlation between K and ϵ. Figure 4 pictures the importance of the choice of experimental points in order to have a confidence region within limits as narrow as possible.

To sum up we emphasize that relations (10) give criteria for choosing the optimal conditions in the determination of K and ϵ by spectrophotometric measurements. The conclusions are not contradictory with those of other authors,[6–8] but the variable G, which

†See ref. 22, page 255.

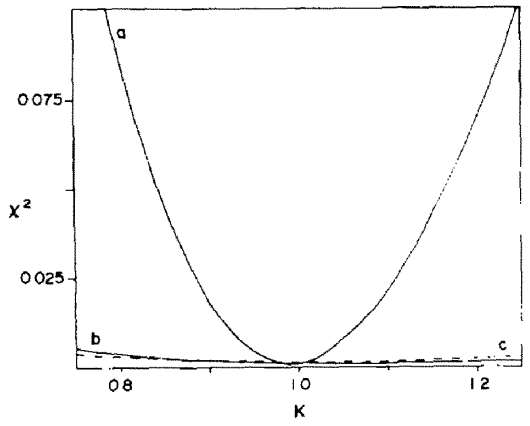

Fig. 2. Minimum values of X^2 vs K, analytically computed,[12] for the three cases *a*, *b* and *c*. this figure shows how the steepness of the different curves, and hence S_k, depends only on the G range which is the same for *b* and *c* (this is true because of the particular G distribution chosen).

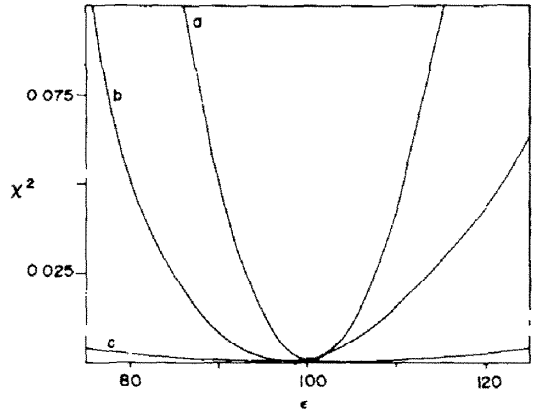

Fig. 3. Minimum values of X^2 vs ϵ, numerically computed, for the three cases *a*, *b* and *c*. This figure shows how the steepness of the curves, and therefore S_ϵ, depends not only on the G range (which is the same for *b* and *c*) but also on the actual G values.

we consider the real independent variable of our system, seems to be more suitable for a correct analysis of the problem. We remark that a non-proper choice of the experimental points, even with a low experimental error, can lead to meaningless K and ϵ estimates.

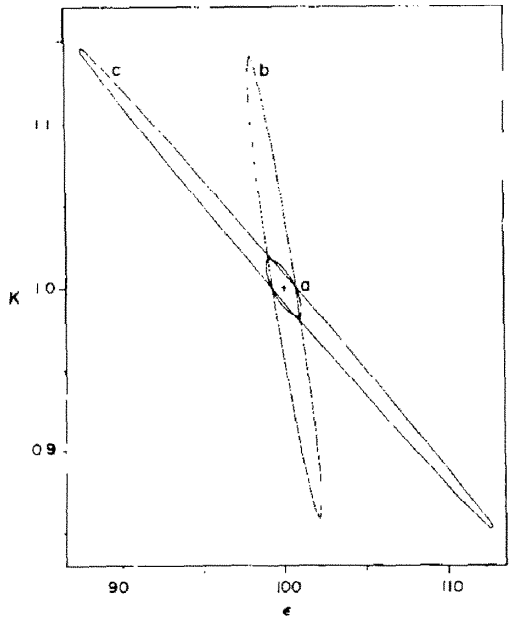

Fig. 4. The ellipses are 90% confidence regions for simultaneous estimates of K and ϵ, relative to the three cases *a*, *b* and *c*. These ellipses have been obtained (by 12) using $S = 0.01$, $K_0 = 1$, $\epsilon_0 = 100$ and the C_{ij} values of Table 3. They clearly show how much the precision of the K and ϵ estimates depends on the different G distribution.

REFERENCES

[1]H. Benesi and J. H. Hildebrand, *J. Am. Chem. Soc.* **71**, 2703 (1949).
[2]J. A. A. Ketelaar, C. van der Stolpe, A. Goudsmit and W. Dzcubas, *Rec. Trav. Chim.* **71**, 1104 (1952).
[3]R. L. Scott, *Ibid.* **75**, 787 (1956).
[4]P. R. Hammond, *J. Chem. Soc.* 479 (1964).
[5]N. J. Rose and R. S. Drago, *J. Am. Chem. Soc.* **81**, 6138 (1959).
[6]W. B. Person, *Ibid.* **87**, 167 (1965).
[7]D. A. Derenleau, *Ibid.* **91**, 4044 (1969).
[8]R. A. La Budde and M. Tamres, *J. Phys. Chem.* **74**, 4009 (1970).
[9]S. D. Christian, E. H. Lane and F. Gasland, *Ibid.* **78**, 557 (1974).
[10]W. C. Coburn and E. Grunwald, *J. Am. Chem. Soc.* **80**, 1318 (1958).
[11]L. G. Sillén, *Acta Chem. Scand.* **16**, 159 (1962).
[12]K. Conrow, G. D. Johnson and R. E. Bowen, *J. Am. Chem. Soc.* **86**, 1025 (1964).
[13]W. E. Wentworth, W. Hirsch and E. Chen, *J. Phys. Chem.* **71**, 218 (1967).
[14]D. R. Rosseinsky and H. Kellawi, *J. Chem. Soc.* A. 1207 (1969).
[15]P. G. Farrel and Phi-Nga Ngo, *J. Phys. Chem.* **77**, 2545 (1973).
[16]W. J. Jones and B. Musulin, *J. Mol. Spectros.* **40**, 424 (1971).
[17]R. Foster and C. A. Fyfe, *Progress in NMR Spectroscopy* (Edited by J. W. Emsley, J. Feeney, L. H. Sutcliffe), Vol. 4, Chap. 1. Pergamon Press, Oxford (1969).
[18]S. D. Ross and M. M. Labes, *J. Am. Chem. Soc.* **76**, 79 (1957).
[19]P. H. Emslie, R. Foster, C. A. Fyfe and I. Horman, *Tetrahedron* **21**, 2843 (1965).
[20]M. Tamres, *J. Phys. Chem.* **65**, 654 (1960).
[21]F. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences.* McGraw-Hill, New York (1969).
[22]O. L. Davies and P. L. Goldsmith, *Statistical Methods in Research and Production.* Oliver & Boyd, Edinburgh (1972).
[23]N. Draper and H. Smith, *Applied Regression Analysis.* Wiley, New York (1966).