Non-parametric Fitting of Nonlinear Equations to Experimental Data without Use of Initial Guessing: A *Basic* Computer Program

Emmanuel M. Papamichael[1*], Nickolaos P. Evmiridis[2] and Constaninos Potosis[1]

[1]Sector of Organic Chemistry and Biochemistry, [2]Sector of Analytical and Inorganic Chemistry, Department of Chemistry, University of Ioannina, 45110 Ioannina, Greece

ABSTRACT

In this work, we present a non-parametric method, and the appropriate computer program, for fitting nonlinear multiparametric equations to experimental data. Our method is followed by computation of confidence limits of the parameter estimates. Its performance has been tested on several multiparametric equations, common in the fields of Biochemistry and Biotechnology, and it is a multiparametric expansion of the concept proposed by others for equations having more than two parameters. Good parameter estimates were obtained without a previous knowledge of initial parameter guessing values, and the proposed computer program converges rapidly, in all cases examined within this work.

Key words: Non-parametric methods, computer program, fitting of nonlinear equations.

INTRODUCTION

A relatively high incidence of outliers is observed during measurements with sensors; some of them are due to the sensing device, others due to the chemistry of the process and others due to sampling procedures. The outliers are a real problem when few replicates are provided; this is a common practice in most kinetic determinations.

This outliers problem becomes a real nuisance when one needs to fit such a series of experimental data to a nonlinear multiparametric equation, that is always the case in Biochemistry, where best parameter estimates are required. The problem becomes more acute when good initial guessing values of the parameter estimates are required.

To overcome this situation different criteria of closeness of fit and/or different kind of fitting algorithms can be used; rules for rejecting outliers can, also, help but they are ineffective in several cases (Anscombe 1960). Alternatively, non-parametric methods have been developed (Eisenthal & Cornish-Bowden, 1974a) where the normality, the uniform variance and other requirements are replaced by an assumption of equally probable positive and negative errors. The latter is not always correct, but it is better than the assumptions made for the parametric methods.

In this report, we present a non-parametric method of fitting nonlinear multiparametric equations to experimental data, the corresponding computer program, and its statistical treatment. Initial guessing values for the parameters of the multiparametric equation under consideration are not required. Our method is an extension of that developed earlier (Eisenthal & Cornish-Bowden, 1974a), and it can be applied to model equations having more than two parameters.

PRINCIPLES

Function Transformation: In many cases the response of a monitored process as it is the rate of a biochemical reaction, is described by a nonlinear multiparametric model equation of the form:

* Author for correspondence

$$\mathbf{y = f(x; a,b,c, . . ,p)} \qquad (1),$$

where **x** and **y** are the independent variable and dependent response, respectively and **a, b, c, . . , p** the parameters. In practice the data x_i,y_i are obtained experimentally. Therefore, each experimental data point (x_i,y_i) should be considered as the known, while the equation parameters as the unknowns to be determined.

In principle, equation (1) can be transformed to the form of a: (i) line equation, (ii) plane equation, (iii) hyperplane equation, for a two or three or more parameters equation, respectively. In the general case the hyperplane equation has the form:

$$\frac{\mathbf{P_1}}{\xi_1} + \frac{\mathbf{P_2}}{\xi_2} + \ldots . + \frac{\mathbf{P_p}}{\xi_p} = \mathbf{1} \qquad (2),$$

where $\mathbf{P = \{P_i\}}$ ($\mathbf{i = 1, 2, . . .,p}$) is the vector of the axial components (unknown terms) of any hyperplane point in a hyperspace, and ξ_i ($\mathbf{i = 1, 2, . . .,p}$) is the intersection of a hyperplane with the $\mathbf{i^{th}}$ axis (known terms). Consequently, a hyperplane is defined for each experimental data point (x_i,y_i) within the given hyperspace.

If errorless data points are applied to a known multiparametric equation the estimated parameter values **a, b, c, . , p** are the same for all data points and therefore the intersection of all defined hyperplanes is a **point** whose coordinates correspond to the true parameter values.

However, due to random experimental errors in measurements the hyperplanes are intersected in several ways, by **two**, by **three**, . , by **p**, while only the intersection by **p** will give **points** in a space of **p** dimensions.

The number of intersections: For equations with **p** parameters, and **n** experimental data points, the maximum total number of hyperplane intersections is given by the formula:

$$\mathbf{N}_{max} = \binom{\mathbf{n}}{\mathbf{2}} + \binom{\mathbf{n}}{\mathbf{3}} + \cdots + \binom{\mathbf{n}}{\mathbf{p}} \qquad (3),$$

where $\binom{\mathbf{n}}{\mathbf{p}} = \frac{\mathbf{n!}}{\mathbf{(n-p)!\ p!}}$ equals to the number

of intersections of **n** hyperplanes by **p** (Binomial Coefficients, **n>p**).

The coordinates of all $\binom{\mathbf{n}}{\mathbf{p}}$ hyperplane intersection **points** that correspond to parameter estimates are collected in **p** columns, and sorted in respect to their arithmetical values. The median value of each column is chosen as the best estimate of the corresponding parameter **a, b, c, . . .,p**.

The error structure: For an infinite number of independent observations (x_i,y_i), errors $\boldsymbol{e_i}$ are supposed to be equally probable either positive or negative. Correspondingly, the true parameter values ***a***, ***b***, , ***p***, are the coordinates of a **point** lying below or above of each hyperplane defined by each (x_i,y_i) data point. This supports our choice to define the median values as the estimates of ***a***, ***b***, , ***p***.

The sample median is regarded as a more reliable estimate of the population average value than the sample mean, and alternative approaches to this mater could be certainly found (Cornish-Bowden & Eisenthal, 1974b). This is outside of the purpose of this manuscript. Herein, we would like to put the idea of non-parametric curve fitting of the multiparametric nonlinear equations; at least, good initial guessing values of the parameters were estimated. The latter being sometimes as the Ancient Greek proverb "The begining (is) the Half of Everything" (Platonis opera, Leges).

An example:

As an example it will be used a three parameter equation, which represents the calibration curve of chemiluminescence generated from oxidized pyrogallol by periodate (Papamichael & Evmiridis, 1988), the following:

$$y = \frac{ax^2}{bx + c\sqrt{x} + 1} \qquad (4).$$

Having a sufficient number **n** of experimental points we will compute estimates of the true values of parameters **a**, **b** and **c** without the knowledge of initial guessing values. For each experimental observation (x_i,y_i), equation (1) could be written as:

$$y_i = \frac{ax_i^2}{bx_i + c\sqrt{x_i} + 1} \qquad (5),$$

which is transformed to:

$$a\frac{x_i^2}{y_i} - bx_i - c\sqrt{x_i} = 1 \qquad (6),$$

for $y_i \neq 0$. Equation (6) falls into a general form of plane Equation (7),

$$\frac{P_1}{\xi_1} + \frac{P_2}{\xi_2} + \frac{P_3}{\xi_3} = 1 \qquad (7)$$

where $P_1 = a$, $P_2 = b$, $P_3 = c$, and

$\xi_1 = \frac{y_i}{x_i^2}$, $\xi_2 = -\frac{1}{x_i^2}$, $\xi_3 = -\frac{1}{\sqrt{x_i}}$ provided that

$\frac{y_i}{x_i^2} \neq \frac{1}{x_i^2} \neq \frac{1}{\sqrt{x_i}}$ ($y_i \neq x_i$ & $x_i \neq 1$) for **i =1,2, . . ,n**.

Due to experimental errors in measurements of the response, equation (5) can be written as:

$$y_i = \frac{ax_i^2}{bx_i + c\sqrt{x_i} + 1} + e_i \qquad (8),$$

where ***a***, ***b***, and ***c*** are the true but unknown values of the **a**, **b**, and **c** parameters, x_i are the errorless values of the independent variable, and e_i is the difference between observed and true response values independently where errors are confined. The total number of intersections should be $\binom{n}{2} + \binom{n}{3}$ but only $\binom{n}{p} = \frac{n!}{(n-3)!\,3!}$ are considered as intersection points.

All mentioned herein, are referred to transformable equations to a suitable form of a hyperplane equation. On the other hand any equation can generally be transformed by arbitrary replacements and proper restrictions.

Equation (9a) can be transformed if an integer value is given, temporarily, to the parameter **d** (i.e. d=1, or d=2), and equation (9b) by making more than one transformations

CONFIDENCE REGIONS OF THE PARAMETER ESTIMATES

Number of regions: In a finite experiment with **n** data points, **n** hyperplanes can be drawn according to equation (2) which divide the hyperspace into 2^n regions at the most. No one hyperplane could be parallel to another and/or do not pass through the origin. These restrictions are in line with the concept of generating hyperplanes from experimental data points that include random errors e_i. This is illustrated in **Fig.1** for **n=10** and **p=3**.

Relative positions of regions and hyperplanes: According to theory there should not one single point of intersection of all **n** hyperplanes defined by the (x_i,y_i) data points, due to random experimental errors e_i (Eq, 8). On the other hand, hyperplanes as they intersected each other divide hyperspace in regions.

Thus, each hyperplane associated with a specific experimental point is either above (e_i positive) or below (e_i negative) of the point the coordinates of which are the true parameter values ***a***, ***b***, ***c***, . , ***p***. That point is, obviously, within these regions and it can be used to define the confidence limits of the parameter estimates (Cornish-Bowden & Eisenthal, 1974b).

Designation of regions: Each e_i error will be either positive(+) or negative(-) and it is associated to a specific region. Therefore, the total number of e_i will be **n**, i.e. e_1, e_2, , e_n. Furthermore, if plus(+) and minus(-) signs of the errors e_i are replaced by **1** and **0** respectively, the designation of the hyperspace regions becomes numerical. Thereafter, regions can be labeled either in a binary or in a decimal form. For example, a region in **Fig. 1**, labeled **24** in decimal form, corresponds to binary 0000011000 for **n**=10 data points or to binary 11000 for **n**=5 data points. In **Fig. 2** a region labeled **696**, in decimal form, corresponds to binary 1010111000.

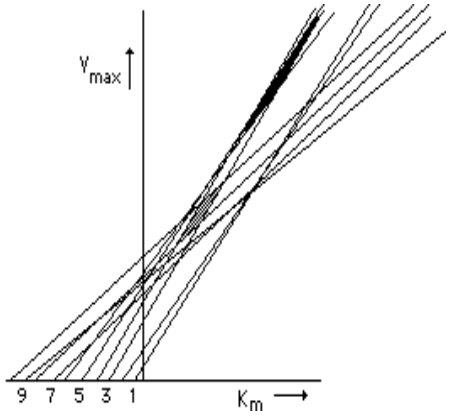

Figure 1: Intersection of ten lines referred to the Michaelis-Menten equation; data were from our laboratory. Dark-shadowed is region **24** (0000011000), and shadowed is region **696** (1010111000).

Permutations: If each e_i value has a median expectation of zero, and all e_i-values are independent, then all possible permutations of signs among e_i values are equally likely. Since the possible permutations of **n** signs is 2^n the probability for each permutation to occur is 2^{-n}. Confidence regions, as just defined, are rigorous but they extend to infinity and include estimates of the true parameter values that must be rejected as absurd.

Table 1 illustrates the way of calculation of Binomial Coefficients and Probabilities with different number of positive and/or negative signs of **n** errors e_i from the 2^n possible permutations. Calculations are given for an equation with three parameters having **n=10**, and for zero positive and ten negative signs, for one positive and nine negative signs, and so on up to ten positive and zero negative signs.

From Table 1 we conclude that there is a probability of about 25% for five positive and five negative signs, moreover there is a greater probability for four to six positive and six to four negative signs. This result provides a confidence region of about 66% for ***a***, ***b***, and ***c*** within experiment comprising those regions that predict about equal number of positive and/or negative signs of the e_i values.

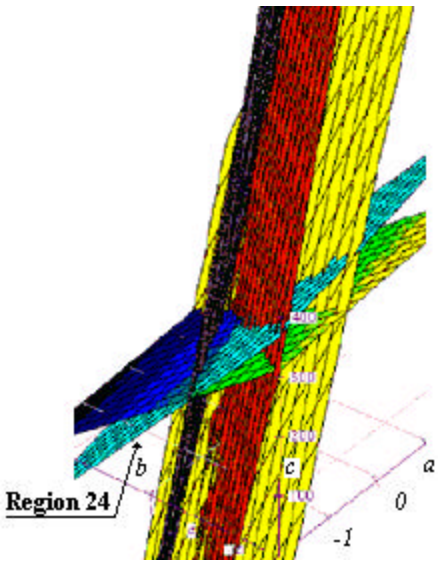

Figure 2: Intersection of five planes (P) referred to equation (4). P1-Cyan, P2-Blue, P3-White (appeared as shadowed), P4-Orange, and P5-Yelow. Region **24** (11000) is indicated in the graph; region **12** (01100) is in the right side between cyan, blue, and orange planes.

Runs: Another way to define confidence regions is that of considering the number of runs, which is the number of separate strings of **0**s and **1**s, in the binary notation, of a confidence region.

Any objection raised on defining confidence regions with the concept of permutations, can be removed by considering the number of runs of positive and negative signs in the series of errors, instead of the total number of positive and negative signs (Cornish-Bowden & Eisenthal, 1974b).

Table 1: Probabilities for + or - signs of errors, for a three parameters equation, and ten experimental points.

Number of Binomial Coefficients		Probability
$\binom{10}{0}$	= 1 (0!=1)	$\frac{1}{2^{10}}$ % = 0.1%
$\binom{10}{1}$	= 10	$\frac{10}{2^{10}}$ % = 1.0%
$\binom{10}{2}$	= 45	$\frac{45}{2^{10}}$ % = 4.4%
$\binom{10}{3}$	= 120	$\frac{120}{2^{10}}$ % = 11.7%
$\binom{10}{4}$	= 210	$\frac{210}{2^{10}}$ % = 20.5%
$\binom{10}{5}$	= 252	$\frac{252}{2^{10}}$ % = 24.6%
$\binom{10}{6}$	= 210	$\frac{210}{2^{10}}$ % = 20.5%
$\binom{10}{7}$	= 120	$\frac{120}{2^{10}}$ % = 11.7%
$\binom{10}{8}$	= 45	$\frac{45}{2^{10}}$ % = 4.4%
$\binom{10}{9}$	= 10	$\frac{10}{2^{10}}$ % = 1.0%
$\binom{10}{10}$	= 1	$\frac{1}{2^{10}}$ % = 0.1%
SUMS	1024	100.0 %

For a region **24** corresponds a sequence of signs - - - - - + + - - - containing three runs (for **n=10**) whereas for a region **696** corresponds a sequence of signs + - + - + + + - - - containing six runs. The number of runs in random sequence of binary digits obey to Binomial Distribution and for **n** digits they are $2\binom{n-1}{m-1}$ permutations with **m** runs. The probability that there are **m** runs in **n** digits is then $\frac{\binom{n-1}{m-1}}{2^{n-1}}$.

In Table 2 is illustrated the calculation of the probabilities for positive and/or negative signs using the same example as in Table 1. A probability of about 50% is calculated for five or six positive (or negative) signs, whereas the probability for four to seven positive (or negative) signs is over than 80%. Conclusions about confidence regions are about the same, using both methods, the latter being more convenient and useful since may give enclosed confidence regions, and relatively small in extent.

CALCULATIONS

The median estimates of the parameters **a**, **b**, **c**, . , **p** are well calculated by solving successive systems of **p** equations, in **p** unknowns. For the example of equation (4), and for $\mathbf{y} \neq 0$ we may write:

$$a\frac{x_i^2}{y_i} - bx_i - c\sqrt{x_i} = 1 \qquad (10),$$

$$a\frac{x_j^2}{y_j} - bx_j - c\sqrt{x_j} = 1 \qquad (11),$$

$$a\frac{x_k^2}{y_k} - bx_k - c\sqrt{x_k} = 1 \qquad (12),$$

where **i, j, k =1**, **2**, **3**, . ., **n**.

Table 2: Probabilities for + or - signs of errors, for a three parameters equation, and ten experimental points.

Fractions of Binomial Coefficients	Permutations and Probabilities
$\frac{\binom{9}{0}}{2^9} = \frac{1}{2^9}$ **(0!=1)**	$\frac{1}{2^9}2^{10} = 2$ **0.19 %**
$\frac{\binom{9}{1}}{2^9} = \frac{9}{2^9}$	$\frac{9}{2^9}2^{10} = 18$ **1.76 %**
$\frac{\binom{9}{2}}{2^9} = \frac{36}{2^9}$	$\frac{36}{2^9}2^{10} = 72$ **7.03 %**
$\frac{\binom{9}{3}}{2^9} = \frac{84}{2^9}$	$\frac{84}{2^9}2^{10} = 168$ **16.41 %**
$\frac{\binom{9}{4}}{2^9} = \frac{126}{2^9}$	$\frac{126}{2^9}2^{10} = 252$ **24.61 %**
$\frac{126}{2^9}$	**252** **24.61%**
$\frac{84}{2^9}$	**168** **16.41%**
$\frac{36}{2^9}$	**72** **7.03%**
$\frac{9}{2^9}$	**18** **1.76 %**
$\frac{1}{2^9}$	**2** **0.19 %**
SUMS	**1024** **100.00 %**

When it happens to be **k=j** or **k=i** or **j=i** then the respective counter is increased by one and we proceed so that to have always **i ≠ j ≠ k**. Another constraint for such a system should be $\frac{x^2}{y} \neq -x \neq -\sqrt{x}$ (for all counters). These non-equalities are important for the presented method. The only constraint for this technique is the equation under consideration to be transformable in a form like that of equation (2).

A variety of methods could be chosen to solve a system of **p** equations in **p** unknown; in this work we preferred to use the Cholesky's or Croutãs method (Shoup, 1983). In the example, equation (10) could be solved to one unknown term, successively, and written as:

$$a = \frac{y_i}{x_i^2} + b\frac{y_i}{x_i} + c\frac{y_i}{\sqrt{x_i}} \qquad (13a)$$

$$b = -\frac{1}{x_i} + a\frac{x_i}{y_i} - c\frac{1}{\sqrt{x_i}} \qquad (13b)$$

$$c = -\frac{1}{\sqrt{x_i}} + a\frac{x_i\sqrt{x_i}}{y_i} - b\sqrt{x_i} \qquad (13c).$$

The above equations represent, one plane in a space of three dimensions which intersects (i) **Z**-axis at the point $\frac{y_i}{x_i^2}$, (ii) **Y**-axis at the point $-\frac{1}{x_i}$ and (iii) **X**-axis at the point $-\frac{1}{\sqrt{x_i}}$. Consequently, intersections by three-planes will be defined by the planes of equations (11) and (12), of which the coordinates of their intersection points give suitable estimates of the **a**, **b**, and **c** parameters.

THE PROGRAM

The program is written in *BASIC* and has been developed under ***ZBASIC*** compiler for Macintosh computers. The user have to edit several program lines before "**RUN**ning" it, as follows:

(a) **Input** the number of (a) data pairs and (b) parameters, of the used equation (constants **ndp** and **npr**).
(b) **Select** the binomial coefficient (binco), from a suitable **TABLE**, according to the data inputted just above, and assign that value to the appropriate constant. Edit **DIM**, and **CLEAR** statements.
(c) **Edit** program lines that describe the system of the linear simultaneous equations, which should correspond to the transformations of the used nonlinear equation; they found within multiple **FOR - NEXT** nested loops.
(d) **Add** as many as **FOR - NEXT** nested loops as it is the number of the parameters of the used nonlinear equation; **add**, also, the appropriate **IF - THEN** statements within nested loops, and edit the end of the last **IF - THEN** statement.
(e) Finally, **add** as many as **[dx( )=...]** assignments, of the loop counters, as it is the number of the parameters of the used equation, run the program and collect results.

RESULTS AND DISCUSSION

The presented method gave excellent results when applied to several multiparametric equations, common in Chemistry and Biochemistry; such as:

$$y = \frac{ax^2}{bx + c\sqrt{x} + 1} \qquad \text{(i);}$$

it is described thoroughly in this manuscript (Papamichael & Evmiridis, 1988; Papamichael & Evmiridis, 1991) as equation (4).

$$y = \frac{x(1-x)}{a+bx+cx^2} \qquad \text{(ii);}$$

it describes the rate of chlorinating of 1,2-dichloroethane to 1,1,2-trichloroethane, and 1,1,1,2- and 1,1,2,2-tetrachloroethanes (Kafarof 1976).

$$y = \frac{ax + bx^3}{1 + cx + dx^5} \qquad \text{(iii);}$$

it describes the rate of substrate depletion by an enzyme that displays non-Michaelis-Menten kinetics.

$$y = \frac{a + bx^2}{1 + cx + dx^2 + ex^3} \qquad \text{(iv);}$$

it is similar that the above one, and describes a different situation of non-Michaelis-Menten kinetic behavior.

$$y = \frac{x(1-x)}{ax + bx^2} \qquad \text{(v);}$$

it describes the same procedure as in (ii), by using two parameters, for a more convenient curve fitting.

Table 3. Illustration of the proposed method. Parameter Estimates from fitting of used Equations to Experimental and/or Simulated data.

Eq.	Parameter Estimates from Non-Parametric Fitting		Parameter Estimates from Ordinary Fitting	
(i)	a =	106.32	a =	100.78
	b =	0.80	b =	0.76
	c =	-1.58	c =	-1.53
(ii)	a =	0.07	a =	0.08
	b =	1.31	b =	1.30
	c =	-0.81	c =	-0.60
(iii)	a =	53.07	a =	53.15
	b =	0.75	b =	0.68
	c =	4.70	c =	4.71
	d =	9.44	d =	9.43
(iv)	a =	38.99	a =	39.00
	b =	11.47	b =	11.41
	c =	-0.16	c =	-0.16
	d =	0.58	d =	0.58
	e =	0.19	e =	0.19
(v)	a =	0.13	a =	0.12
	b =	0.90	b =	0.96

Experimental (x_i,y_i) data pairs were used for both the ordinary and the non-parametric fitting of the above equations (i) to (iii); for

equations (iv) and (v) were used simulated data produced accordingly (Papamichael & Evmiridis, 1991). For the ordinary fitting were used either commercial curve fitting packages (UltraFit, 1991), or self-written programs (Papamichael & Evmiridis, 1988).

The results are summarized in Table 3. Estimates of the parameters of all equations were found very close to their true values. The program was converged after few seconds in most cases; in case of equation (iv) of Table 3 the program converged after 1.6 h, most probably due to the large number of parameters and data points. In difficult situations, when the number of data points are relatively few compared to the number of parameters of the used equation, the program was given parameter estimates which can be used as good initial guessings for other methods.

Objections could be raised on the efficiency and/or the usefulness of the proposed method and program. At a first glance, (a) the method is not applicable for certain nonlinear equations e.g. the Gompertz equation $\mathbf{y = ae^{-be^{-cx}}}$ (Ratkowsky, 1983), or (b) all linear forms of nonlinear multiparametric equations can be fitted to any set of data points by multivariate linear regression methods without need of initial guessed values for their parameters.

Any equation can be generally transformed to a suitable form like equation (2) by arbitrary replacements and proper restrictions including the Gompertz one (Ratkowsky, 1983). On the other hand, in linear regression we make assumptions on which we based in order to accept that the minimum variance estimators are unbiased and normally distributed. However, to do so in nonlinear situations we need a very large number of data which are unavailable as far as it is concern the fields of Biochemistry and/or Biotechnology, in most cases (Ratkowsky, 1983).

Based on the above we can conclude that the efficiency, the statistical robustness, and the usefulness of the proposed method and program have been verified.

REMARKS

The Program Listing is available on request. It is written under Z-BASIC for Macintosh; however, there is a version of the Z-BASIC compiler for IBM compatibles. Alternatively, authors could help on transformation of the Program Listing under any available compiler.

RESUMO

Neste trabalho, apresentamos um método não-paramétrico, e o programa de computação apropriado, para ajustar equações multiparamétricas não lineares para dados experimentais. Nosso método é seguido pelo cálculo dos limites de confiança dos parâmetros estimados. Seu desempenho tem sido testado em várias equações multiparamétricas, comuns nos campos da Bioquímica e Biotecnologia, e é uma expansão multiparamétrica do conceito proposto por outros, para equações que tenham mais de dois parâmetros. Obtivemos estimativas de parâmetros confiáveis sem um conhecimento prévio de parâmetros iniciais de valores esperados, e o programa de computação proposto converge rapidamente, em todos os casos examinados dentro deste trabalho.

REFERENCES

Anscombe .F.J. (1960), Rejection of outliers *Technometrics*, **2**, 123-147.

Eisenthal R., Cornish-Bowden A. (1974a), The direct linear plot: A new graphical procedure, *Biochem J.*, **139**, 715-720.

Eisenthal R., Cornish-Bowden A. (1974b), Statistical considerations in the estimation of enzyme kinetic parameters by the direct

linear plot and other methods, *Biochem J.*, **139**, 721-730.

Kafarov V., (1976), Cybernetic Methods in Chemistry & Chemical Engineering, MIR Publishers, Moscow, pp.376-379.

Papamichael E. M., Evmiridis N. P., (1988), Program based on the pattern search method: application to periodate determination using FIA analysis and chemiluminescence detection, *TrA.C.*, **7**, 366-370.

Papamichael E. M., Evmiridis N. P., (1991), Investigation of the error stucture of the calibration curve for, periodate determination by FIA analysis and chemiluminescence detection, *C.I.L.S.*, **7**, 39-47.

Plotonis opera, Leges (Plato's works laws), **6,** 753.

Shoup T. E, (1983), Numerical methods for the Personal Computer, PRENTICE HALL, INC. Englewood Clifts, New Jersey, pp.45-57.

UltraFit, (1991), The non-Linear curve fitting package, BIOSOFT, Cambridge U.K., pp. 5-48.

Received: February 15, 1999;
Revised: April 18, 1999;
Accepted June 06, 1999.