Revista Mexicana de Física 20 (1971) 239 - 252

STRUCTURE ANALYSIS OF PHYSICAL DATA
I. GENERAL THEORY

C. Marmasse

Facultad de Ciencias,
Laboratorio de Biofísica, UNAM

(Recibido: agosto 3, 1971)

ABSTRACT: The problem of measuring the degree of concordance between a mathematical structure and a set of experimental data is examined. It is solved after an iterative fitting of the data to a linearized form. The method permits an easy evaluation of various fiducial boundaries. In particular it is shown that the degree of concordance sought can be measured by means of only one parameter. The latter can also be used to test the homogeneity of a set of data with respect to a given mathematical structure.

I. INTRODUCTION

A new type of problem concerning the numerical exploitation of experimental data has become more and more acute: the problem of the selection, based on quantitative and objective cirteria, between various possible mathematical structures in order to represent in physical terms a given set of experimental data. In particular, it is the fundamental and classical problem

of physical research; however a new dimension is added because in many cases only a relatively small number of points is available. This in turn requires the development of a well- adapted fitting procedure.

The need for such a method is best demonstrated by a numerical example. The equation which describes the inhibition of an enzyme by an excess of its substrate has been given by Haldane[1]* as

$$\frac{v_{max}}{v} = 1 + \frac{K_M}{s} + \frac{s}{K_{SS}} \qquad (1)$$

where v is the initial velocity of the reaction, s the concentration of the substrate and v_{max}, K_M and K_{SS} three parameters. The basic problem is to determine these three parameters. Extrapolation and trial and error methods which have been extensively used have proved of doubtful value. On the other hand, an approximate method, known as the linear method was described earlier[3]. One can also rewrite equation (1) as

$$\frac{1}{v} = \frac{1}{v_{max}} + \frac{s}{K_{SS} v_{max}} + \frac{K_M}{s v_{max}} \qquad (2)$$

and interpret (2) as a bilinear regression of the dependent variable $1/v$ onto the two independent variables s and $1/s$. In these conditions, a fitting of the experimental data to (2) by means of a conventional least squares method will furnish estimates of the three parameters. But Haldane's equation can also be written as

$$\frac{s}{v} = \frac{K_M}{v_{max}} + \frac{s}{v_{max}} + \frac{s^2}{v_{max} K_{SS}}. \qquad (3)$$

This relation can be interpreted as a quadratic regression of the variable s/v onto the independent variable s; therefore a fitting of the experimental data to relation (3) by means of a conventional least-square method will also furnish estimates of the three parameters.

* The notation used here is that of Bray and White (2)

TABLE 1. Hydrolysis of acetylcholine chloride by acetylcholinesterase from *Helix* blood. These data, borrowed from Augustinsson[4], are expressed in arbitrary units and have been corrected for non-enzymatic hydrolysis.

s	v
14.8	120
49.3	356
148	419
493	413
1479	324
4928	194

TABLE 2. Estimates of the parameters obtained by various methods of fitting the data of Table 1. The results are given in the arbitrary units used by Augustinsson.[4]

Fitting	v_{max}	K_M	K_{SS}
Linear method	636	61.6	2100
To equation (2)	467	18.3	3514
To equation (3)	606	56.7	2281

The result of the fittings of the data of Table 1 are presented in Table 2. The existence of very large discrepancies between the estimates illustrates the need for careful and more elaborate analysis of the experi-

mental data. This conclusion is reinforced when one notes* that in the case at hand, the "correct" values are

$$v_{max} = 543 \qquad K_M = 2430 \qquad K_{SS} = 34.5$$

It is therefore clear that the straightforward use of a least square procedure is unsatisfactory, since it leads only to a dead-lock; a more sophisticated approach is necessary.

First of all, one needs a fitting procedure which is both efficient and unbiased; the more so as the sets of experimental data available often contain but a relatively small number of points. Under these circumstances, one cannot afford to let uncertainties, not accounted for, creep into the calculations, as the effect of these uncertainties could be unexpectedly magnified at a later step of the analysis.

This is not, however, the end of the matter. It is extremely important to obtain fiducial limits for the computed values of the parameters. In practice, this second aspect is crucial. The central problem is to assess the degree of concordance (compatibility) between a given set of experimental data and a given mathematical structure. A second problem, most commonly encountered, is assessing the likelihood that two sets of experimental data come from the same population.

Finally, a third condition should also be met, as far as possible: it is clearly desirable to design the statistical procedure sought so that it can be used easily, routinely and as automatically as possible without sacrificing in any way the efficiency and preserving the absence of bias. In practice, this means that this method should be easy to implement on a digital computer and in such a way as to require but little computing time.

The first and third criteria point to an interative linearization procedure, weighted if necessary. It will be shown that this method has, when the weights can be calculated*, the great advantage of permitting a quick estimation of those values of the kinetic parameters which are statistically the best, and of leading to the easy construction of fiducial limits. In addition, it permits the characterization, by means of one parameter only, of the degree of concordance between a given set of experimental data and a given mathematical structure.

---

* See paper II of this series and ref. 10.

* For a set of examples in enzyme kinetics see ref. 5.

2. FITTING PROCEDURE

As was pointed out in the last section, the general solution of the problem at hand is to be found in an iterative linearization procedure. The origin of such a technique can be traced back to Gauss and some of its properties are well known. However as its use in the particular context brings to light quite a number of new problems, it shall be explicitely treated. It will be noted that in accordance with the third criteria the whole analysis is developed here in such a way that it can be easily programmed.

Let us consider the function y of the r variables x

$$y = f(x_1, x_2, \ldots, x_r; K_1, K_2, \ldots, K_N)$$

and let us assume that we already know a set of approximate values ${}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N$, for the set of parameters $\{K_l\}(l = 1, 2, \ldots, N)$. The set of the "best" values in the statistical sense will be denoted by $\{\tilde{K}\}(l = 1, 2, \ldots, N)$. These values statistically satisfy

$$y = f(x_1, x_2, \ldots, x_r; \tilde{K}_1, \tilde{K}_2, \ldots, \tilde{K}_N) \quad . \qquad (4)$$

One can always define increments $\Delta K_l (l = 1, 2, \ldots, N)$ by means of the N relations

$$\Delta K_l + {}^{\lambda}K_l = \tilde{K}_l \qquad l = 1, 2, \ldots, N \; . \qquad (5)$$

We have, through substitution of (5) into (4),

$$y = f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1 + \Delta K_1, {}^{\lambda}K_2 + \Delta K_2, \ldots, {}^{\lambda}K_N + \Delta K_N) \; . \qquad (4)$$

It will be noted that when the increments $\Delta K_l (l = 1, 2, \ldots, N)$ are zero, or do not significantly differ from zero, the parameters ${}^{\lambda}K_l (l = 1, 2, \ldots, N)$ become equal to or do not significantly differ from the parameters $\tilde{K}_l (l = 1, 2, \ldots, N)$.

Under proper conditions of convergence, equation (4) can be expanded, within a domain $\mathbb{D}$, as

$$y = f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N) + \sum_{l=1}^{N} \frac{\partial f}{\partial {}^{\lambda}K_l} \Delta K_l + \sum_{l,m=1}^{N} O(\Delta K_l \Delta K_m) . \quad (6)$$

Equation (6) rewritten as

$$y = f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N) + \sum_{l=1}^{N} \frac{\partial f}{\partial {}^{\lambda}K_l} \Delta K_l \quad (7)$$

can be interpreted as a multilinear regression of the dependent variable y on the $(N + 1)$ independent variables

$$f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N) \text{ and } \frac{\partial f}{\partial {}^{\lambda}K_l} \quad l = 1, 2, \ldots, N .$$

By fitting the set of experimental points to relation (6), values of the increments ΔK_l $(l = 1, 2, \ldots, N)$ can be obtained and, in turn, better starting values ${}^{\lambda+1}K_l$ $(l = 1, 2, \ldots, N)$ can be obtained by

$${}^{\lambda+1}K_l = {}^{\lambda}K_l + \Delta K_l \quad l = 1, 2, \ldots, N .$$

The computation can be iterated and the increments ΔK_l $l = 1, 2, \ldots, N$, can be reduced to zero or nearly so, within the domain $\mathbb{D}$.

However, the regression (6) is not a conventional regression in the sense that the coefficient of the first independent variable, *i. e.* $f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N)$ must be equal to 1. In order to remove

this statistical constraint, we shall use the new dependent variable

$$z = y_{exp} - f(x_1, x_2, \ldots, x_r; {}^{\lambda}K_1, {}^{\lambda}K_2, \ldots, {}^{\lambda}K_N)$$

where y_{exp} stands for the observed value of y.

In the following developments we shall always suppose, except when otherwise specified, that the structure of the function f is such that it has a constant – or nearly constant – variance. Transformation techniques suitable for achieving this result will be indicated later.

In these conditions, one expects z to have a nearly Gaussian distribution; and the closer to a Gaussian distribution, the closer the set of the parameters $\{K_l\}$.

The new system to be solved is then

$$z = \sum_{i=1}^{N} \frac{\partial f}{\partial K_i} \Delta K_i$$

$$z \equiv y_{exp} - f \ .$$

Let us define* a real symmetric matrix T and a vector $\boldsymbol{p}$ by means of

$$t_{ij} = \sum_{\alpha=1}^{N} \left(\frac{\partial f}{\partial K_i}\right)_{\alpha} \left(\frac{\partial f}{\partial K_j}\right)_{\alpha}$$

$$p_i = \sum_{\alpha=1}^{N} z_{\alpha} \left(\frac{\partial f}{\partial K_i}\right)_{\alpha} , \quad i, j = 1, N$$

where n is the number of experimental points.

Let be Δ the vector of general element ΔK_l. The normal equations take the

---

* The reader unfamiliar with regression analysis may consult the works by Williams[6] and/or Kendall[7].

$$T\Delta = p$$

and therefore

$$\Delta = T^{-1}p \ .$$

3. FIDUCIAL LIMITS

The total sum of squares is

$$\sum_{\alpha=1}^{n} z_\alpha^2 \ .$$

As the regression hyperplane is mathematically compelled to pass through the origin, there are only N degrees of freddom. On the other hand the sum of squares accounted for by the regression is $\Delta \cdot p$. Therefore the residual mean squares s^2 (variance) is estimated by

$$(n-N)\, s^2 = \sum_{\alpha=1}^{n} z_\alpha^2 - \Delta \cdot p \ .$$

One will note that the variance so determined is the sample variance; within the framework of the maximum likelihood theory, the best estimate of the population variance is $\mathcal{B} s^2$ where $\mathcal{B}$ is the Bessel correction factor

$$\mathcal{B} = \frac{n-N}{n-N-1} \ .$$

The variance of the increment ΔK_l is $s^2 t^{ll}$ where t^{ij} is the general element of T^{-1}. As the $\{\Delta K_l\}$ are normally distributed, fiducial limits can be obtained as

$$\Delta K_l \pm st\sqrt{t^{ll}} \qquad (7)$$

where t is the value of the Student t-distribution at the selected level of confidence.

These estimates of the variance of the various increments can be used to establish statistically the concordance of the computed values K_i, $i = 1, 2, \ldots, N$, with the experimental data. It is sufficient to establish the statistical lack of significance of the increments ΔK_i, $i = 1, 2, \ldots, N$. Such a test is exact because the null hypothesis defines a set of values K_i; under these circumstances, the regression is a N-linear regression on the *a priori* given functions $\partial f / \partial K_i$. In practice, however, the test described hereafter is more efficient and more convenient.

The limits described by relations (7) are not simultaneous. The probability that these limits are simultaneously reached is smaller than the probability corresponding to the selected t point.

In order to eliminate this "loss of information", one can set up a statistical test for the simultaneous departure from zero of the N increments, following Williams method.[6] Let us consider the quadratic form

$$Q_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta K_i \, \Delta K_j \, t_{ij} \ ,$$

which is a form with N degrees of freedom. Therefore, the ratio $Q_0/(Ns^2)$ is distributed as F with N and n–N degrees of freedom*. It follows that the fiducial boundary at a given probability level is given by

$$Q_0 = Ns^2F \ .$$

Clearly this relation also defines the fiducial boundary for the set of the $\{K_i\}$ at the probability level. The physical meaning of this fiducial boundary can be described as follows:

Given a set of approximative values K_i, $i = 1, N$ a computed set of increments ΔK_i – that is to say also, a set of values $K_{i,\,\text{improved}} = K_i + \Delta K_i$ – is compatible at a given probability level with the experimental data, with respect to a given mathematical structure, if they satisfy the inequality

$$Q_0 \leqslant Ns^2F \ .$$

* The reason why there are n-N degrees of freedom in this formula and not (n-N-1) as in the classical formula, is that the regression is mathematically (as opposed to statistically) constrained to pass through the origin.

We shall now derive from the form Q_0 a new form Q_x ; this permits the assessment of the compatibility, at a certain probability level, of a set of experimental data with a given mathematical structure by means of one parameter endowed with physical meaning. Let us first consider the quadratic form defined by

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} (\Delta K_i - \alpha_i)(\Delta K_j - \alpha_j)\, t_{ij}$$

where the α_i represent *a priori* values of the corresponding increments ΔK_i . An argument similar in all points to the one just touched on shows that Q is distributed as F with N and $n-N$ degrees of freedom.

Let us now consider the influence of a simultaneous variation of the increments by a certain fraction x of the accepted values of the corresponding kinetic parameters. Then

$$\alpha_i = xK_i$$

and the form Q becomes

$$Q_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} (\Delta K_i - xK_i)(\Delta K_j - xK_j)\, t_{ij} \ .$$

Now the iteration is stopped when the $|\Delta K_i / K_i|$ are very small. Therefore

$$\Delta K_i < xK_i$$

when x is not too small, which is the usual case, and in these conditions the form Q is not very different from the form Q_x defined by

$$Q_x = x^2 \sum_{i=1}^{N} \sum_{j=1}^{N} K_i K_j t_{ij} \ .$$

Let us put

$$Q_K \equiv \sum_{i=1}^{N} \sum_{j=1}^{N} K_i K_j t_{ij} \quad .$$

Here Q_K is a constant for a given set of experimental data and a given mathematical structure. The fiducial boundaries are then defined by

$$x^2 Q_K = NFs^2 \quad .$$

Let x_1 be the value obtained by giving to F the value one in the preceeding equation. We have

$$x_1 = s\sqrt{N/Q_K}$$

Now the only meaningful values of F are those equal to or greater than one. It follows that a simultaneous departure of all the parameters by a fractional amount less than or equal to x_1 is compatible at any level of confidence with the set of experimental data and for the mathematical structure considered. Or, in other words, the values of the parameters cannot be trusted to have a simultaneous relative precision better than x_1.

In a general way, simultaneous fiducial limits at a given level of confidence can be calculated by

$$\zeta_i (1 \pm x) = K_i (1 \pm x_1 \sqrt{F}) \quad .$$

This formula clearly shows that the knowledge of the parameter x_1 is sufficient to calculate simultaneous fiducial limits at any probability level that is to say that x_1 is a *practical parameter* which characterizes the degree of compatibility of a set of experimental data with a given mathematical structure.

4. DISCUSSION

One will note that no restriction* was placed on the function f, save that of having homogeneous variance.

Formally, this restriction is not much of a burden. Let us denote the variance, if it exists, of a function g by var (g); then the new variable $g/\mathrm{var}(g)$ has an homogeneous variance. It follows that if one substitutes $z/\mathrm{var}(f)$ and $[1/\mathrm{var}(f)]\,\partial f/\partial K_i$ for z and $\partial f/\partial K_i$, respectively, the results of the preceding sections are most general. In order to analyze a physical phenomenon, considerations of simplicity and convenience in accordance with the third requirement laid down in section 1, may serve as guidelines for the selection of an appropriate structure of f. The estimation of $\mathrm{var}(f)$ requires a careful analysis of the physical nature of the measurements carried out. This is to say that the weights of the regression are determined by the physical method used to obtain the data analyzed; (for an example concerning the analysis of spectrophotometric data, see ref. 8). This point which is, in particular, most essential for a correct interpretation of the x_1 statistic will be discussed at greater length in the following papers of this series.

The parameter** x_1 can be used, in particular, to study the homogeneity of a set of experimental data with respect to a given mathematical structure. This comes from the fact that this parameter can be interpreted as a measure of the goodness of fit; in this condition, if one decreases the importance attached to a given point by means of an appropriate weighting factor, the parameter x_1 will tend to decrease, all other things being equal, if the contribution of this point to the variance is important, *i. e.* if it is much different from the others. A scanning of the experimental data by means of a systematic depression of each point in turn, will thus enable one to identify anomalous data (provided that variance of z is constant).

The characterization by means of the parameter x_1 of the degree of compatibility between one set of experimental data and a given structure is very efficient and convenient when no *a priori* information is available. If, on the other hand, possible values for one or several of the parameters were to be tested, the form Q would have to be used. In particular, this form is advantageous to test if two sets of experimental data are distinguishable or

---

* In addition to implicit assumptions of continuity for example.

** Note that the parameter x_1 can also be calculated by means of the relation $Q_1(x_1) = ns^2$. This permits to relax somewhat the conditions $|\Delta K_i/K_i| \ll 1$ when they lead to a prohibitive amount of computing time.

not at a given probability level and with respect to a given mathematical structure. Attributing indices a and b to the quantities related to the two groups of data, we have explicitly

$$Q_{\text{test}} = \sum_{i,j=1}^{N} [\Delta K_{a,i} - (K_{b,i} - K_{a,i})] [\Delta K_{a,j} - (K_{b,j} - K_{a,j})] t_{a,ij}$$

Let there be a lower limit α of the level of confidence at which the relation

$$F_{N,n-N,\alpha} \geqslant \frac{Q_{\text{test}}}{Ns^2}$$

holds. Then a measure of the degree of compatibility expressed in % is given by

$$\delta = 100\,(1 - \alpha)$$

Coming back to the analysis of one set of data with respect to a given mathematical structure, one will note* that one can take into account the possible stiffness of the mathematical structure studied in some parameters by expressing the α_i as

$$\alpha_i = x\rho_i K_i$$

where ρ_i is a measure of the "desired precision" of the parameter K_i.

Then the form Q (and the form Q_K if the $|\Delta K_i / K_i|$ are sufficiently small) becomes a function of the ρ_i, *i.e.* of quantities endowed with a concrete meaning. This would enable one to study the effects of the relative precision of various selected parameters on the fiducial boundary for the set of the parameters.

---

* The author is indebted to the referee for this suggestion and would like to take this opportunity to thank him for many comments which have helped clarify this paper.

Finally one will note that, as the fitting procedure described above furnishes unbiased and efficient estimates of the parameters, simple statistical studies on the set of the ultimate $\{z_i\}$ can give valuable results. The combinatorial formulae of the theory of runs[9] are especially convenient. Such an analysis can be applied to the sequence of the signs of the $\{z_i\}$[10]: a repartition markedly different from a random distribution is indicative of a systematic deviation. In certain cases, this type of analysis can successfully complement a scanning analysis performed with the x_1-parameter.

REFERENCES

1. Haldane, J.B., *Enzymes*, Longmans, Green and Co., London 1930, p. 84
2. Bray, H.G. and White K., *Kinetics and Thermodynamics in Biochemistry*, Academic Press, New York 1957, p. 225
3. Marmasse, C., Biochim. Biphys. Acta, 77 (1963) 530.
4. Augustinsson, K.B., Acta Physiol. Scand., (1948), Suppl. 15,52
5. Marmasse, C., Lecture Notes, Latin American School of Physics. México 1971
6. Williams, E.J., *Regression analysis*, Wiley, New York 1959
7. Kendall, M.G., *The advanced theory of statistics*. Charles Griffin and Co., London; Vol. I, 1945; Vol. 2, 1950
8. Marmasse, C., Appl. Opt., (1965), 4, 1932
9. Feller, W., *An introduction to probability theory and its application*, Wiley, New York, 2nd. ed. 1960
10. Marmasse, C., Rev. Mex. Fís., 19 (1970) 146.

RESUMEN

Se analiza el problema de la medida del grado de concordancia entre una estructura matemática y un conjunto de datos experimentales. Es solucionado después de un ajuste iterativo de los datos a una forma linearizada. El método permite evaluar varias fronteras fiduciales. En particular, el grado de compatibilidad se puede medir por medio de un solo parámetro, que también puede ser empleado para probar la homogeneidad de un conjunto de datos respecto a una estructura matemática dada.