Biophysical Chemistry, 36 (1990) 85–89
Elsevier

BIOCHE 01442

Point perturbation analysis of experimental data

II. The statistics of relaxation length values of pseudo-random errors

Enrico Di Cera, Francesco Andreasi Bassi and Giuseppe Arcovito

Istituto di Fisica, Universita' Cattolica, Largo F. Vito 1, 00168 Roma, Italy

Received 25 September 1989
Revised manuscript received 28 December 1989
Accepted 28 December 1989

Data analysis; Monte Carlo simulation; Residuals

The statistics of relaxation lengths for pseudo-random deviates as determined by point perturbation analysis (E. Di Cera, F. Andreasi Bassi and G. Arcovito, Biophys. Chem. 34 (1989)239), has been constructed by a Monte Carlo study. The values of the relaxation length, L, approximately follow a Gamma distribution. The results allow for a statistical estimation of relaxation profiles and provide a test for randomness of residuals which is more accurate than other standard procedures.

1. Introduction

We have recently proposed a new method of checking for systematic errors of experimental data based on 'point perturbation analysis' (PPA) [1]. The gist of this method stems from construction of a perturbation matrix from the stretch of residuals derived from linear or nonlinear least squares of experimental data. The perturbation matrix is then processed by standard techniques of Fourier analysis [2,3] to yield the periodogram and the autocorrelation function of each experimental point. Integration of the autocorrelation function yields a 'relaxation length', L, for each residual that represents a distance of relaxation for the correlation of each residual [1].

Preliminary investigation has shown that a value of $L > 1$ for a given experimental point should be considered indicative of significant correlation and hence nonrandomness [1]. A key question of course arises as to the statistics associated with relaxation length values. In the present study, we wish to report on a Monte Carlo investigation of the expected distribution of relaxation length values for pseudo-random deviates, in order to obtain a statistics for L in the absence of systematic errors. The results provide a sound statistical basis for the quantitative interpretation of relaxation profiles obtained by PPA.

Correspondence address: E. Di Cera, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110, U.S.A.

2. Methods

The PPA method is discussed in detail elsewhere [1] and is briefly outlined here solely to make this paper self-contained.

Consider the n-dimensional vector of experimental data $\boldsymbol{y}$ and the fitting function $F(\boldsymbol{x},\boldsymbol{p})$, where $\boldsymbol{x}$ is the independent variable and $\boldsymbol{p}$ the vector of fitting parameters, and the associated functional [4]

$$\Phi = [\,\boldsymbol{y} - F(\boldsymbol{x},\boldsymbol{p})\,]^T\,[\,\boldsymbol{y} - F(\boldsymbol{x},\boldsymbol{p})\,] = \boldsymbol{r}^T\boldsymbol{r} = \sum_j r_j^2 \qquad (1)$$

where $\boldsymbol{r}$ is the vector of residuals and summation * is taken over the n experimental data points. The minimum (best-fit) value of ϕ is given by

$$\Phi^* = \boldsymbol{r}^{*T}\boldsymbol{r}^* = \sum_j r_j^{*2} \tag{2}$$

and is a function of n residuals. Another functional of the form given in eq. 1 can be minimized after dropping a particular data point, say y_k. This functional is given by

$$\Phi_k'^* = \boldsymbol{r}_k^{*T}\boldsymbol{r}_k^* = \sum_{j\neq k} r_{kj}^{*2} \tag{3}$$

and is the sum of $n-1$ residuals of the form r_{kj}^*, where the subscript k indicates that the residual of point j has been calculated by dropping point k. The term r_{kk}^* gives the residual of the point being dropped, i.e., the distance of y_k from the new best-fit value of F obtained when point k is weighted zero in the minimization procedure. By adding r_{kk}^{*2} to eq. 3 one has the important stability relation [1]

$$\Phi_k'^* + r_{kk}^{*2} = \sum_j r_{kj}^{*2} = \Phi_k^* \geq \Phi^* \tag{4}$$

which states that Φ_k^* is the functional Φ^* 'perturbed' by dropping point k and is always greater than the unperturbed functional, independent of the particular point being dropped.

An $n \times n$ perturbation matrix is then constructed with elements

$$\delta r_{jk}^* = r_{jk}^* - r_k^* \tag{5}$$

where δr_{jk}^* gives the displacement of the k-th residual from its 'equilibrium' position, r_k^*, due to the perturbation induced by dropping the j-th point [1]. In the absence of systematic errors one expects, for any k, the n changes $\delta r_{1k}^*, \delta r_{2k}^*, \ldots \delta r_{nk}^*$ to be random. This can be established by processing each column of the perturbation matrix by standard techniques of Fourier analysis in order to extract harmonic components indicative of nonrandom behavior [1,2]. For each point k the discrete Fourier transform of frequency $\tau = j/n$ is given by

$$f_k(\tau) = \sum_j \delta r_{jk}^* e^{-i[2\pi(j-1)\tau]} \tag{6}$$

and the corresponding periodogram is

$$I_k(\tau) = n f_k(\tau) f_k^0(\tau)/2 \tag{7}$$

where f^0 is the complex conjugate of f. The autocorrelation function $C_k(\alpha)$ for each point is given by the inverse Fourier transform of $I_k(\tau)$ and is an even function computed in the lag range $\pm q$ ($q = n/2$ for n even; $q = (n-1)/2$ for n odd). The function $C_k(\alpha)$ is tapered with Bartlett's weights [1,2]

$$W(\alpha) = \begin{cases} 1 - \alpha/q & \text{for } \alpha \leq q \\ 0 & \text{for } \alpha > q \end{cases} \tag{8}$$

and the relaxation length for each point, $L(k)$, is finally obtained as the integral

$$L(k) = \int_{-q}^{q} C_k(\alpha) W(\alpha)\, d\alpha \tag{9}$$

The relaxation length for a point subject to random noise in the range $\pm\infty$ must be zero, as a consequence of the fluctuation-dissipation theorem [5]. In practice, since one only deals with finite values of q and the autocorrelation function is tapered according to eq. 8, then the integral above does not return a value of zero for random deviates, but rather yields a relaxation profile close to the zero baseline [1]. The key question arising is to establish the exact distribution of relaxation length values in the case of random deviates and to construct a statistics thereof. This would allow for a quantitative interpretation of the relaxation profile, especially in the case of L values well above the zero baseline. Such a problem can best be addressed by a Monte Carlo study.

Random noise perturbations of a given experimental point were generated as δr^* values using pseudo-random deviates of the form [6]

$$\delta r_j^* = \sqrt{-2\ln(\mathrm{RND}_1)} \cdot \cos(2\pi \mathrm{RND}_2) \tag{10}$$

where RND_1 and RND_2 are two random numbers such that $0 < \mathrm{RND} < 1$ and $j = 1, 2 \ldots n$. Ini-

* We adopt the same notation as in ref. 1: matrices and vectors are denoted by bold and summations are always taken from 1 to n unless otherwise noted.

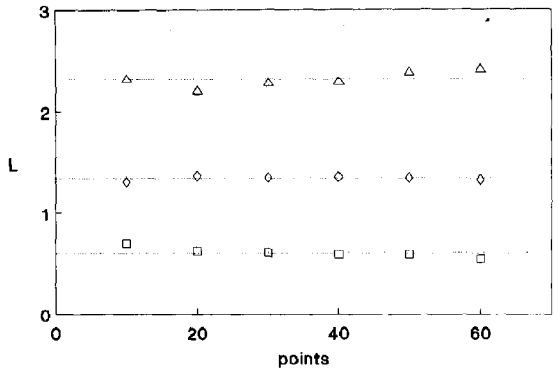

Fig. 1. Relaxation values obtained as a function of residual points at the cutoffs of one (□), two (◇) and three (△) standard deviations from the Monte Carlo simulations discussed in the text. Horizontal lines depict the average value of each cutoff and correspond to L values of 0.60 (□), 1.34 (◇) and 2.32 (△).

tially, 10^5 perturbations of variable length were simulated with n ranging from 10 to 60 in order to assess the dependence of the statistics on the size of the perturbation matrix. Each sequence composed by n δr^* values, generated according to eq. 10, was processed using eqs. 6–9 to yield a corresponding value of L. The distribution of all 10^5 L values was statistically analysed for the different values of n. The L values corresponding to one, two and three standard deviations of the distributions obtained are shown in fig. 1 from which it is clearly observed that these cutoffs are practically independent of the value of n in the range studied. The same conclusion was reached after generating δr^* values according to the logistic difference equation [7], i.e.,

$$\delta r_j^* = x_j \tag{11}$$

with

$$x_j = a x_{j-1}(1 - x_{j-1}) \tag{12}$$

where $0 < x < 1$, and $x_0 = 0.35$. The equation above yields a periodogram characteristic of random noise for $a = 4$, which corresponds to a dynamical regime of complete chaos [8,9].

3. Results

The distribution of relaxation length values for the case $n = 10$ is shown in fig. 2. The underlying statistics is not Gaussian, but rather follows quite closely the Gamma distribution

$$f(L) = L^{\alpha} e^{-\beta L} \tag{13}$$

with $\alpha = 1.06 \pm 0.02$ and $\beta = 3.50 \pm 0.05$. The probability that L lies in the range from 0 to X is given by

$$P(L \le X) = \left\{ \beta^{\alpha+1} / \Gamma(\alpha+1) \right\} \int_0^X L^{\alpha} e^{-\beta L} dL \tag{14}$$

where Γ is the Gamma function. The probability that L exceeds X is then

$$P(L > X) = 1 - P(L \le X) \tag{15}$$

From the relations above, one readily calculates the statistical confidence of a value of $L = X$ not being random as $P(L \le X)$. The probability that $L \ge 1$ is only 14.3% for pseudo-random deviates and hence a value of $L > 1$ is indicative of systematic errors with at least 85.7% confidence. In the case of $n > 10$ the agreement between the distribution of L values and the Gamma distribu-

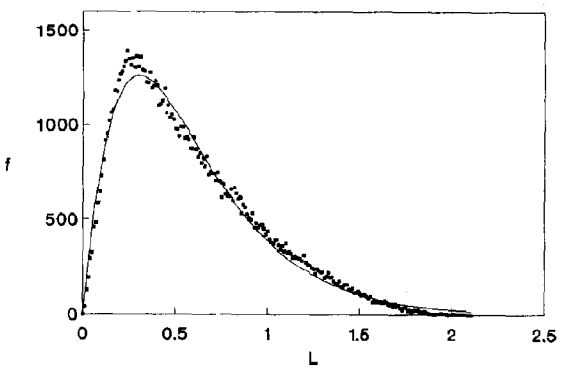

Fig. 2. Distribution of relaxation length values for 10^5 stretches of 10 pseudo-random deviates obtained by Monte Carlo simulation as discussed in the text. The data are shown as absolute frequencies (or number of observations), f, versus L. The L values were cumulated in discrete intervals of 0.01 L units. The continuous line was drawn according to the Gamma distribution $AL^{\alpha}e^{-\beta L}$ with best-fit values: $A = (1.29 \pm 0.06) \times 10^4$; $\alpha = 1.06 \pm 0.02$; $\beta = 3.50 \pm 0.05$.

tion is not as good as in the case $n = 10$, and therefore the statistical confidence associated with any value of L must be computed directly from the cumulative distributions obtained by Monte Carlo simulation. Interestingly, the cutoff values for one, two and three standard deviations appear to be independent of n (see fig. 1). These values can thus be used as 'invariant' statistical quantities for practical applications of PPA.

4. Discussion

The results of the foregoing analysis allow for a quantitative interpretation of relaxation profiles obtained by PPA. In a previous paper [1], the value of $L = 1$ was taken as cutoff for assessing nonrandom behavior of a residual. In view of the distribution of L values obtained for pseudo-random errors we can conclude that this cutoff is extremely accurate. In fact, it gives a confidence of 85.7% in considering a value of $L = 1$ to be 'suspect' and hence indicative of nonrandom behavior. It might be of interest to note that a 68.3% (one standard deviation) confidence is associated with a value of L as low as 0.60.

The relaxation profile of shear viscosity measurements of a 2,6-lutidine-water mixture [1] is shown in fig. 3, along with that obtained from viscosity values simulated with pseudo-random errors. The horizontal dotted lines depict confidence cutoffs of one (68.3%), two (95.5%) and three (99.7%) standard deviations. These lines facilitate the quantitative assessment of nonrandom (or random) behavior for each experimental data point. One sees that, in the case of viscosity measurements simulated with superimposed random errors, the relaxation profile is very low with points below the cutoff of one standard deviation. In the case of experimental measurements of shear viscosity the relaxation profile shows points well above the cutoff of two standard deviations, which strongly indicates the presence of systematic errors in agreement with the conclusions drawn in a previous study [1].

Finally, we wish to stress the higher discriminative power of PPA as compared to that of other standard tests of residuals [2–4]. The relaxation

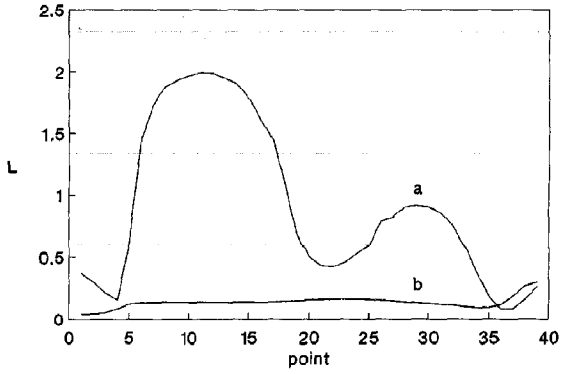

Fig. 3. Relaxation profiles of: (a) experimental determinations of the temperature dependence of shear viscosity of the 2,6-lutidine/water mixture [1], and (b) viscosity values simulated in the same temperature range with superimposed pseudo-random errors. The experimental determinations are given in table 1 of ref. 1. Dotted lines depict the values of L expected for truly random deviates at the cutoffs of one (68.3%), two (95.5%) and three (99.7%) standard deviations (see fig. 1).

profiles shown in fig. 3 clearly point out the dramatic difference between experimental determinations and the data set simulated with superimposed pseudo-random errors. The results of the standard autocovariance test of residuals [2–4] for the two sets are shown in fig. 4. No significant departure from random behavior can be seen in both cases and the two profiles look very similar. Another standard test of residuals is obtained by plotting the residuals on probability paper [3,4].

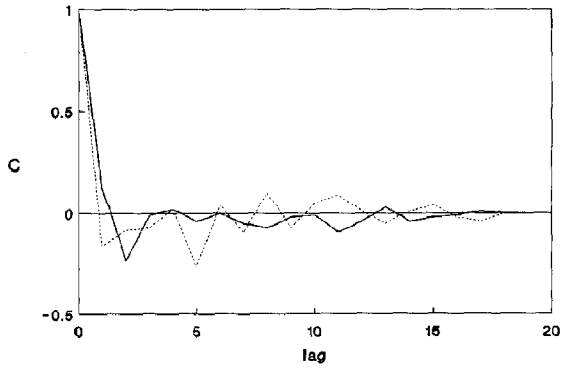

Fig. 4. Autocovariance test of residuals [2–4] for the same experimental (continuous line) and simulated (dotted line) viscosity measurements shown in fig. 3. The autocovariance is very low in both cases and indicates no appreciable departure from random behavior.

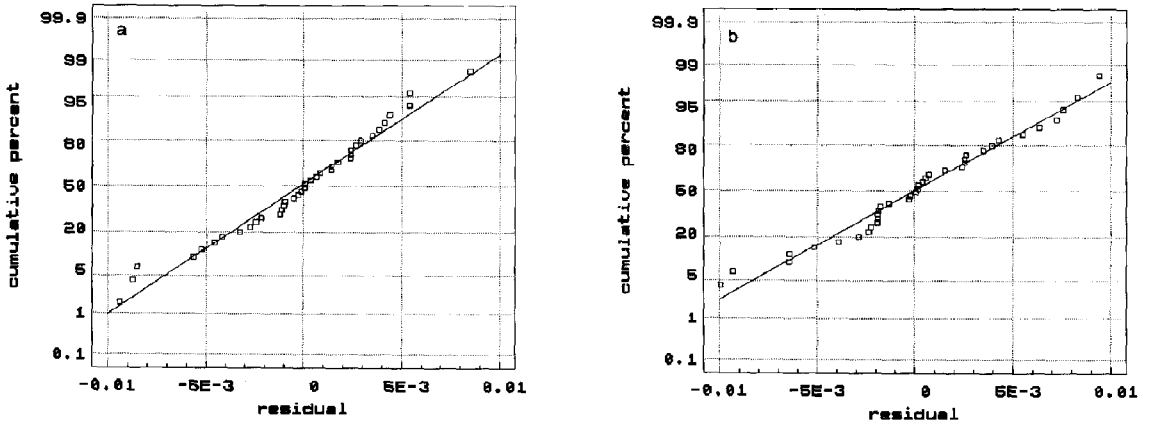

Fig. 5. Probability test of residuals [3,4] for the same experimental (a) and simulated (b) viscosity measurements shown in fig. 3. Note the particular scale given on the ordinates for such a test. The residuals are scattered around a straight line in both cases, which implies no appreciable departure from random behavior.

Random residuals in such a plot are uniformly scattered around a straight line. The results of this test for the two cases shown in figs 3 and 4 are depicted in fig. 5. Again, no appreciable difference can be seen between experimental and simulated data sets, and the residuals are scattered around a straight line in both cases. Therefore, from the standard tests of residuals shown in figs 4 and 5 one would conclude that the experimental determinations of shear viscosity of the 2,6-lutidine water mixture [1] contain no appreciable systematic errors, as the residuals behave like those obtained in the case of a simulated data set with superimposed pseudo-random errors. On the other hand, a quite different conclusion can be reached via PPA, as indicated by the relaxation profiles shown in fig. 3. The higher resolution of the PPA method [1] is certainly due to the information gained through point perturbations, by which a stretch of n residuals is expanded into an $n \times n$ perturbation matrix.

Acknowledgements

This work was supported by MPI and CNR.

References

1 E. Di Cera, F. Andreasi Bassi and G. Arcovito, Biophys. Chem. 34 (1989) 239.
2 M.B. Priestley, Spectral analysis and time series (Academic Press, New York, 1981).
3 G.E.P. Box and G.M. Jenkins, Time series analysis (Holden-Day, Oakland, CA, 1976).
4 Y. Bard, Nonlinear parameter estimation (Academic Press New York, 1974).
5 S.R. de Groot and P. Mazur, Non-equilibrium thermodynamics (Dover, New York, 1984).
6 W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical recipes (Cambridge University Press, New York, 1986).
7 R.M. May, Nature 261 (1976) 459.
8 B. Bunow and G.H. Weiss, Math. Biosci. 47 (1979) 221.
9 E. Di Cera and J. Wyman, J. Mol. Liq. 41 (1989) 33.