Information Content and Retrieval in Solution Scattering Studies. I. Degrees of Freedom and Data Reduction

BY DANIEL TAUPIN

Laboratoire de Physique des Solides, Centre Universitaire, 91405 – Orsay, France

AND VITTORIO LUZZATI

Centre de Génétique Moléculaire, CNRS, 91190 – Gif sur Yvette, France

(Received 21 May 1981; accepted 15 December 1981)

Abstract

As a consequence of several technological developments, the speed and sensitivity of solution X-ray (and neutron) scattering studies have recently risen by factors which may be as large as 10^5; formal problems – like assessing the information content of the data and retrieving that information in structural terms – have thus become of immediate practical interest. The general problem of the information content of scattering experiments is discussed; its mathematical expression is derived, which depends on both the experimental data (observed values and estimated accuracy) and the a priori stochastic assumptions on the structure of the sample. The practical application of these notions to solution scattering studies involves several steps, three of which – choice of the degrees of freedom, data reduction and error analysis – are dealt with in this work. The first step is to specify the minimal number of independent parameters necessary and sufficient to describe the whole of the scattering properties of the system. Whenever the solute particles are of finite dimensions the entire scattering curve is defined by its values at a one-dimensional lattice; if, moreover, the asymptotic trend of the scattering curves is known, then the degrees of freedom are the ideal intensities at a finite number of points plus a small number of parameters describing the asymptotic trend. It is also possible to include among the degrees of freedom a few subsidiary parameters like the normalization factors. The next step is, starting from a composite set of data, to determine the most probable numerical value of each degree of freedom and to evaluate its range of uncertainty. This is discussed within the framework of variable-contrast studies, assuming that the invariant-volume hypothesis is fulfilled. An algorithm is formulated which treats all the experimental observations and determines simultaneously all the degrees of freedom and the error matrix. The algorithm also allows one to introduce additional linear constraints on the degrees of freedom. As an example, the algorithm is applied to solution X-ray scattering data recorded with a low-density serum lipoprotein. The determination of the maximal chord of the particle – an important parameter in the informational analysis – turns out to be rather tricky.

1. Introduction

X-ray (and neutron) scattering are well established tools for structural studies of macromolecules in solution [see recent reviews in Pilz, Glatter & Kratky (1979) and Luzzati & Tardieu (1980)]. Yet, until a few years ago, solution X-ray scattering studies suffered from severe technical limitations; much too often, indeed, the reward for cumbersome and time-consuming experiments was poorly informative and disappointingly inaccurate results. These limitations were particularly striking by contrast with other structural techniques – especially X-ray crystallography and electron microscopy – which over the last three decades have become so remarkably powerful and versatile. Then, beginning in the early seventies, a series of technological innovations made a sudden and dramatic impact in solution X-ray scattering studies: these were the introduction of position-sensitive X-ray detectors (Borkowski & Kopp, 1968; Gabriel & Dupont, 1972; Fourme, Bordas & Koch, 1981) and the use of synchrotrons and storage rings as X-ray sources (Stuhrmann, 1978). The overall effect was that in a span of a few years exposure times dropped by more than 10^5. At almost the same time the use of high-flux reactors also transformed neutron scattering into a powerful technique for the structural study of macromolecules in solution (Ibel, 1976).

One of the effects of these technological advances has been a revived interest in some theoretical aspects of solution scattering, previously neglected owing to the lack of practical motivations. More specifically, as the experimental data were piling up at increasing speed in computer memories, problems like assessing the amount of information, retrieving this information...
and translating it into physical terms became vitally urgent. We have tackled these problems in a few recent papers (Luzzati, Tardieu & Aggerbeck, 1979; Luzzati, 1980; Luzzati & Tardieu, 1980; see also Damaschun, Müller & Pürschel, 1968; Moore, 1980; Ramakrishnan & Moore, 1982). Yet, the reference to the concepts of information theory is not quite satisfactory in any of the previous publications, whose scope is in fact restricted to two aspects of the general problem of information content and retrieval, namely specification of the degrees of freedom and data reduction.

In this and in other forthcoming papers, we set ourselves to look into the same general problem from a more rigorous standpoint, borrowing concepts and techniques from probability and information theories. Our purpose is multifold:

1. As in previous papers, we sort out the degrees of freedom of the problem, namely the smallest possible number of independent parameters, necessary and sufficient to describe the whole of the scattering properties of the system.

2. We deal with data reduction, our goal being to formulate a set of algorithms which, starting from a composite set of raw data, determines the most probable values of the degrees of freedom, and evaluates their probability distribution.

3. We strive to measure the information content in precise statistical terms. This step involves not only the values and uncertainties of the degrees of freedom, but also an a priori stochastic physical model of the system.

4. We attempt to retrieve the information in the physical space by reference to the physical model; this step is equivalent to narrowing down the ranges of likelihood of the very parameters which are left at random in the a priori stochastic model. We expect, at this stage, to be in a position to look into some statistical notions of immediate practical interest (accuracy, resolution) so commonly and so loosely used in structural studies based on scattering methods; our purpose is to establish quantitative criteria investing those concepts with clear mathematical sense.

5. We strive to extend the same theoretical treatment to crystallographic studies, with the practical purpose of establishing unified quantitative criteria to assess and compare the intrinsic quality of any particular set of scattering data, should it come from single crystals or from solution, obtained using X-rays or neutrons.

II. Mathematical treatment

II.1. A priori, conditional and joint probabilities

The physical and the mathematical operations involve two categories of variables: $Y$, which specifies the experimental measurements (for example $Y_k$ may be the count recorded at the $k$th channel of the detector), and $X$, which specifies the parameters corresponding to the degrees of freedom of the problem (for example the ideal intensity scattered at one particular scattering angle by a perfect sample and recorded by a perfect diffraction apparatus). Those two categories of variables are not necessarily homogeneous nor are their numbers of elements equal. If a vectorial notation is used,

$$X = \{X_k\}$$ for the parameters (degrees of freedom) 

$$Y = \{Y_k\}$$ for the experimental data. 

The transfer function of the instrument can be expressed by the conditional probability density $q(Y|X)\,dx$ that $Y$ be inside the volume element $dx$ when the value of $X$ is known.

The choice of a stochastic model (for example $N$ atoms located at random inside a volume $V$) specifies the a priori probability $r(X)dx$ that the parameter vector $X$ (for example the ideal intensities at various scattering angles) is in the volume element $dx$. Inversely, one can introduce the a priori probability of $Y$ without any information on $X$: $u(Y)dx$, and the conditional probability of $X$ with $Y$ known: $r(X|Y)dx$.

All these functions can be condensed in a joint probability function $z(X,Y)dx\,dy$. These definitions can be summarized as:

$$t(X)dx$$ a priori probability of $X$, \hspace{1cm} (II-1-2a)

$$u(Y)dy$$ a priori probability of $Y$, \hspace{1cm} (II-1-2b)

$$q(Y|X)dx$$ conditional probability of $Y$ with $X$ known, \hspace{1cm} (II-1-2c)

$$r(X|Y)dy$$ conditional probability of $X$ with $Y$ known, \hspace{1cm} (II-1-2d)

$$z(X,Y)dx\,dy$$ joint probability of $X$ and $Y$. \hspace{1cm} (II-1-2e)

These functions are linked together by the following relationships:

$$z(X,Y) = q(Y|X)r(X) = r(X|Y)u(Y)$$ \hspace{1cm} (II-1-3a)

$$t(X) = \int z(X,Y)dy$$ \hspace{1cm} (II-1-3b)

$$u(Y) = \int z(X,Y)dx$$ \hspace{1cm} (II-1-3c)

$$r(X|Y) = \frac{z(X,Y)}{u(Y)} = \frac{q(Y|X)r(X)}{\int q(Y|X)r(X)\,dx}$$ \hspace{1cm} (II-1-3d)

From the experimental standpoint, we assume that all the components of $Y$ are distorted by an additive non-correlated noise, which is assumed to be Gaussian. Moreover, the most probable value of each $Y_k$ is expected to have an explicit expression $Y_k(x)$. Under these conditions, $q(Y|X)$ can be written as a product of independent Gaussian probability functions, which in turn can be expressed as a multi-dimensional Gaussian law:
The information content associated with the observation of an event whose probability was \( P \) before performing the observation is defined by

\[
H = -\log_2 P. \tag{II-2-1}
\]

In the present case, the 'event' is a set of observations leading to a new probability law. The probability of such an event can be calculated by reversing to a slightly different statistical problem: what is the probability that one variable \( X \), whose \textit{a priori} distribution is \( t(X) \), is observed with a distribution \( r(X|Y_o) \)? The answer is derived in various classical papers (Kullback, 1968, pp. 3–6). Using previous notations, \( H \) can be written:

\[
H = \int r(X|Y_o) \log_2 \left\{ \frac{r(X|Y_o)}{t(X)} \right\} \, dX. \tag{II-2-2}
\]

This expression of the information content involves several elements:

(a) The choice of the degrees of freedom (the components of \( X \)).

(b) The determination of the most probable values of \( X \).

(c) The estimation of \( q(Y|X) \) which is needed only in (II-2-2) [see remark (II.1.c)].

(d) The choice of a stochastic physical model, defining the \textit{a priori} probability \( t(X) \).

The elements (a), (b), (c) are discussed below; (d) will be dealt with in a forthcoming paper.

II.3. Degrees of freedom: an application of Shannon's theorem

A typical solution scattering experiment consists in recording the intensity scattered by the solution, by the solvent and by the instrument (sample cell, slits, \textit{etc.}) and evaluating the instrumental distortions (mainly collimation and polychromatism). The result, usually digital, is a set of intensities associated with a number of channels. It is clear that the intensities recorded at the different channels may well be not independent, especially when the number of channels is large; it is worthwhile to note in this respect that some of the questionable operations commonly performed on the intensity curves (smoothing, interpolations, extrapolations) are based upon the very presence of such correlations. The aim of this section is to specify the degrees of freedom of the problem, namely the minimum number of independent parameters necessary and sufficient to describe the entire scattering curve (see Damaschun et al., 1968; Müller & Damaschun, 1979; Luzzati et al., 1979; Luzzati, 1980; Moore, 1980).

It is hardly possible to tackle this problem without making some assumption about the structure of the sample. The following conditions define the framework of our treatment (see Luzzati, 1980; Luzzati & Tardieu, 1980):

(a) The sample is an ideal solution of discrete particles.

(b) The particles are all identical.

(c) The particles are 'globular', i.e. none of their dimensions is large with respect to \( (s_{min})^{-1} \), \( s_{min} \) being the lower limit of the interval of \( s \) explored experimentally.

An obvious consequence of the globular shape is that the autocorrelation function \( p(r) \) vanishes beyond the maximal chord of the particle \( D_m \). Since the function \( s_i(s) \) is the Fourier transform of \( r(p(r)) \), then Shannon's theorem (Shannon, 1949) shows that the minimum number of independent parameters necessary and sufficient to define the entire function \( s_i(s) \) are the values of this function at the lattice points \( h \Delta s \), \( \{h=1,2,3,\ldots \infty \} \), with

\[
\Delta s = 1/2D_m. \tag{II-3-1}
\]

As a rule the number of lattice points is infinite and thus no finite set of data suffices to define the function \( i(s) \) completely. The problem becomes manageable if, beyond some point \( s_m \), the form of \( i(s) \) can be expressed as a function of a finite number \( L \) of parameters. In this case the entire function \( i(s) \) is defined by its values at the \( 2D_m s_m \) lattice points in the interval \( 0 < s < s_m \), plus the values of the \( L \) parameters. Thus the number of degrees of freedom becomes

\[
J = 2D_m s_m + L. \tag{II-3-2}
\]

In X-ray scattering studies of systems of biological interest the function

\[
i(s) = A/s^4 + B, \quad s > s_m, \tag{II-3-3}
\]

has been shown to provide a satisfactory description \( i(s) \) beyond \( s \sim (25 \, \text{Å})^{-1} \) (see references in Luzzati & Tardieu, 1980); in this case \( L=2 \).
Using the notation of § II-1, we have:

\[ \{ Y_k \} = \{ Y(S_k) \} \]

the counts recorded at the nominal values \( s = s_k \) (II-3-4)

\[ \{ X_j \} = \{ [w_j, A, B] \} \]

the degrees of freedom

\[ W_h = h \Delta s (i(h \Delta s)) \] (II-3-5a)

\[ \{ X_j \} = \{ [W_j, A, B] \} \]

the degrees of freedom (II-3-5a)

\[ W_n = hAs (i(hAs)). \] (II-3-5b)

The expression of the ideal \( s_i(s) \) calculated for a given set of values \( \{ X_j \} \) takes the form (Luzzati, 1980; Moore, 1980):

\[ h_m \]

The expression of the ideal \( s_i(s) \) takes the form (Luzzati, 1980; Moore, 1980):

\[ h_m \]

The expression of the ideal \( s_i(s) \) takes the form (Luzzati, 1980; Moore, 1980):

\[ h_m \]

The expression of the ideal \( s_i(s) \) takes the form (Luzzati, 1980; Moore, 1980):
where \( i(s, \phi) \) is an ideal intensity curve in the experimental space, without errors and distortions, \( i_{\text{spherical}}(s, \phi) \) is the spherically averaged intensity curve, \( m \) is the number of electrons of one molecule, and where

\[
K = c_{\text{e}} \eta E_{\text{o}} v
\]

is a normalization factor, \( c_{\text{e}} \) being the electron concentration, \( \eta \) the thickness of the sample, \( E_{\text{o}} \) the energy of the incident beam, and \( v \) a physical constant.

The above expression (II-4-1) can be related to the (expected) experimental counts by (see II-3-9)

\[
Y_{\rho}(\phi) = K \{ \int C(s, s_{\phi}) [i_{\text{spherical}}(s, \phi)/m] ds \}.
\]

The normalization factor \( K \) (see II-4-2 and II-4-3) is the product of several experimental parameters, all determined with finite accuracy. The proper way of dealing with those parameters is to introduce one scaling factor \( K_{\rho} \) for each \( \rho \)-th curve (recorded with \( \phi = \phi_{\rho} \)) and to consider its true \( (K_{\rho_{\text{true}}}) \) and measured \( (K_{\rho_{\text{measured}}}) \) values as additional components of the vectors \( X \) and \( Y \) respectively.

Using the notations adopted in §§ II.1 and II.4, we can now summarize the components of the vectors representing the experimental data and the degrees of freedom:

\[
Y_{\rho} = \{ Y_{\rho, \rho}, K_{\rho_{\text{true}}}, K_{\rho_{\text{measured}}} \}
\]

\[
X = \{ [W_{\xi}], [A_{1}], [B_{1}], [K_{\rho}], [K_{\rho_{\text{true}}}, K_{\rho_{\text{measured}}}] \},
\]

where \( \xi = 1, 2, 3 \) stands for subscripts \( \rho, \rho_{\text{v}} \) and \( \text{v} \) respectively. \( Y_{\rho, \rho} \) is the count actually recorded in the \( k \)-th channel and in the \( \mu \)-th sub-experiment (i.e. when \( \rho = \rho_{\mu} \)) and where (see II-3-5b):

\[
W_{\xi, h} = h \Delta s_{\xi} i_{\xi}(h \Delta s)/m.
\]

According to (II-3-10), (II-4-1) and (II-4-2), the expression for \( Y_{\rho}(\phi) \) takes the form:

\[
Y_{\rho}(\phi, X) = K_{\rho}(\phi) \sum_{\xi = 1}^{3} (\rho - \rho_{\xi})^{-1} \left\{ \int_{\Delta s} W_{\xi, h} \Psi(s_{\rho}, h) + \sum_{k = h + 1}^{h_{\mu}} \{ A_{1}(h \Delta s)^3 + B_{1} h \Delta s \} \Psi(s_{\rho}, h) \right\}.
\]

Then, in the vicinity of \( Y_{\rho}(\phi, X_{\text{o}}) \) (II-1-4b) becomes

\[
S^{2}(X) = \sum_{k, \rho} \left( Y_{\rho, \rho} - Y_{\rho}(\phi, X_{\text{o}}) \right)^2 / \sigma_{\rho}(s_{k}, \rho_{\rho}, X_{\text{o}})^2 + \sum_{\rho} \left( K_{\rho} - K_{\rho}(\phi_{\rho}) \right)^2 / \sigma_{K_{\rho}}^{2}(\phi_{\rho}, X_{\text{o}}). \]  

A few comments can be made:

(a) From a formal viewpoint, the quadratic function (II-4-1) of \( \rho \) involving the characteristic functions can be visualized as an empirical interpolation algorithm describing the \( \rho \) dependence of the X-ray scattering experiments over the explored range of solvent densities, without involving assumptions about the physical properties of the sample.

(b) Other experimental determinations can be involved in the equation above: for example, blank scattering (from solvent, cell, slits, etc.), the function \( C(s, s_{\phi}) \) describing collimation and chromatic distortions (see II-3-9), the value of \( s \) corresponding to the \( k \)-th channel. Whenever the expected values of these experimental determinations can be expressed as functions of a few parameters, then these parameters can be treated like the normalization factors \( \{ K_{\rho} \} \) (see II-4-4, II-4-5 and II-4-8), namely as additional components of the vectors \( X \) and \( Y \), associated with additional terms in the sum (II-4-8).

II.5. Most probable values and uncertainties of the degrees of freedom and of related functions

All the knowledge relevant to the degrees of freedom is contained in the probability distribution \( r(X|Y_{o}) \) (see II-1-3d), which involves both the a priori distribution \( t(X) \) - defined by the stochastic model and the a posteriori distribution \( q(Y_{o}|X) \), defined by the experimental observations. Yet \( t(X) \) does not necessarily have to be defined with great accuracy. For example, one often knows beforehand that \( t(X) \) is much flatter than \( q(Y_{o}|X) \) (in other words little is known before the experiment is performed); in this case the expression of \( r(X|Y_{o}) \) involves only \( q(Y_{o}|X) \) (see II-1-3d):

\[
r(X|Y_{o}) = \frac{q(Y_{o}|X)}{\int q(Y_{o}|X) dx}.
\]

This will be assumed hereinafter. The case of \( t(X) \) being a delta function in one or a few directions, and quite flat in all the others, i.e. the case of constraints binding the components of \( X \) together and restricting the actual number of degrees of freedom, will be discussed at the end of this section.

Owing to the form of (II-1-4), the determination of the most probable value \( X_{\text{o}} \) is equivalent to the determination of the minimum of

\[
L(X) = 2 \log q(Y|X) = S^{2}(X) + \sum_{k} \log \sigma_{k}(X).
\]

If the variations of the logarithm are neglected, the minimization of \( L(X) \) reduces to the minimization of the function \( S^{2}(X) \) in the \( X \) space; besides, a usually satisfactory approximation to the evaluation of the uncertainties is based upon a development of \( S^{2}(X) \) into the differences \( \{ X_{k} - X_{o} \} \).

When the function \( Y_{\rho}(\phi) \) is a linear function of \( X \) and when the \( \{ \sigma_{k} \} \) do not depend strongly on \( X \) (this is a more restrictive hypothesis than neglecting their logarithm), then the error sum \( S^{2}(X) \) becomes a classical quadratic function of \( X \), of the form (\( \chi^2 \) is the ‘transposition’ symbol):

\[
S^{2}(X) = X^T \hat{A} X + 2 \hat{B} \cdot X + S^{2}(O),
\]

\( O \) being the origin, and where the matrix \( \hat{A} \) and the
vector $\mathbf{B}$ can be evaluated by summing the second- and first-order derivatives of each term in the sum (II-4-8). The result is the well known least-squares determination of the most probable values, namely the solution of the equation:

$$\hat{A} \mathbf{X}_o = -\mathbf{B}. \quad \text{(II-5-4)}$$

In fact our expression of $S^2(\mathbf{X})$ (II-4-8) is not a linear function of the whole vector $\mathbf{X}$ but only of each of its component $X_i$. In this case the overall minimum of $S^2(\mathbf{X})$ can be determined using the following iterative algorithm:

(i) assume, as a first approximation that $K_{c\alpha} = K_{o\alpha}$; i.e. use experimental values as starting values;
(ii) solve the restricted system (II-5-4) for $\{W_{c,\alpha}\}$;
(iii) minimize $S^2$ successively for the subset $\{K_{c\alpha}\}$; this operation involves a system of equations similar to (II-5-4);
(iv) go back to step (ii) above, with updated values of the $\{K_{c\alpha}\}$, until convergence is satisfactory.

(algorithm II-5-5)

These operations also offer an opportunity to check the statistical assumptions. The first test is the reduced value of $S^2(\mathbf{X}_o)$ at the optimum $\mathbf{X}_o$:

$$S^2_\text{red} = \frac{S^2(\mathbf{X}_o)}{(N-J)} = \left\{ \sum_{k} \left[ Y_{o_k} - Y_{c\alpha}(\mathbf{X}_o) \right]^2 / \sigma^2_{c\alpha}(\mathbf{X}_o) \right\} / (N-J), \quad \text{(II-5-6)}$$

where $N$ and $J$ are respectively the numbers of observations and of degrees of freedom. The second and more stringent test – the $\chi^2$ test – is a comparison between the observed and expected distributions of the deviations between observed and calculated values.* The $\chi^2$ test is usually performed for multiple occurrences of a unique random variable, with a unique probability density; it can easily be extended to the present case of distinct normalized random variables $\{Y_k\}$ and distinct variances $\{\sigma^2_k\}$, by introducing an equivalent random variable:

$$U = (Y_k - Y_{c\alpha}) / \sigma_k. \quad \text{(II-5-7)}$$

the expected distribution $P(U)$ of which is independent of $k$ (see II-1-4). The $\chi^2$ test is performed by dividing the range of variation of $U$ into a number of slots, and determining the frequency $f_i$ of deviations falling into slot $i$. If $P_i$ is the probability that a single reduced deviation falls into slot $i$, the expression of the reduced $\chi^2$ takes the form:

$$\chi^2_{\text{red}} = \frac{1}{N-J} \sum_{i} \left[ f_i - N P_i \right]^2 / N P_i. \quad \text{(II-5-8)}$$

If all the hypotheses are fulfilled, then the values of $S^2$ and of $\chi^2_{\text{red}}$ are expected to be very close to 1. The fact of observing values substantially different from 1 may be due to several causes:

(a) The mathematical model defining $Y_{c\alpha}(\mathbf{X})$ (see II-4-7) is wrong.
(b) The statistical model for the noise is incorrect.
(c) Some unforeseen systematic error is present (strictly speaking this is equivalent to the mathematical model being inadequate).

It is also worth noting that both $S^2$ and $\chi^2$ can be evaluated for a subset of the measurements $\{Y_k\}$; this property can be exploited to locate the source of the discrepancy in the $Y$ space (see § III.1).

Once the most probable vector $\mathbf{X}_o$ is determined, the next step is to obtain at least an approximate expression of $S^2$ around $\mathbf{X}_o$, i.e. the matrix of its second derivatives:

$$\hat{A}_o = \left\{ \frac{\partial^2 S^2}{\partial X_j \partial X_k} \right\}. \quad \text{(II-5-9)}$$

In the simplest case where the vector $\mathbf{Y}$ is a linear function of the vector $\mathbf{X}$, and when the $\{\sigma_k\}$ are or can be considered as independent of $\mathbf{X}$, the matrix $\hat{A}_o$ is identical to $\hat{A}$ (in II-5-3); if any of these conditions is not met, then all the second derivatives of $S^2$ must be calculated.

The uncertainty of a given degree of freedom $X_j$ can now be estimated using the matrix $\hat{A}_o$: when the other terms of $\mathbf{X}$ are fixed, then the uncertainty in $X_j$ takes the form:

$$\sigma^2_{\text{unc}_j} = \frac{1}{\hat{A}_{o,j,j}}. \quad \text{(II-5-10)}$$

In fact the interesting parameter is not the uncertainty when the other degrees of freedom are supposed to be known, but rather the uncertainty in $X_j$, independently of the value taken by $\{X_{k \neq j}\}$, within the probability law $r(\mathbf{X}|\mathbf{Y}_o)$. This parameter can be determined by integrating the function $r(\mathbf{X}|\mathbf{Y}_o)$ with respect to all other components $\{X_{k \neq j}\}$; the result is:

$$\sigma^2_{\text{ind}_j} = \hat{A}^{-1}_{o,j,j}. \quad \text{(II-5-11)}$$

In addition to the degrees of freedom, one is often interested in other functions and parameters whose expressions are linear combinations of the degrees of freedom. The most common examples are:

(a) The ideal intensity $I_1(\mathbf{s}, \rho)$ (II-4-1) interpolated between the Shannon lattice points and at any solvent density by means of (II-4-7) and (II-3-12).

(b) The autocorrelation function $r(\mathbf{s}, \rho)$:

$$r(\mathbf{s}, \rho) = \frac{1}{2} \sum_{\zeta} \left( \hat{\rho}_\zeta - \hat{\rho}_0 \right)^2 \left\{ \sum_{k=1}^{h} \left( W_{c,k} \right)^2 \right\} \sin(2\pi \rho h \Delta s) - h \Delta s B_{\delta} \sin(2\pi \rho h \Delta s) + \sum_{h=\Delta m + 1}^{\Delta} \left[ A_{\zeta}(\Delta s)^3 \right] \sin(2\pi \rho h \Delta s). \quad \text{(II-5-12)}$$

* Unfortunate notations in textbooks (e.g. Bevington, 1969, compare pp. 84 & 101) lead to a frequent confusion between these two tests.
The values of the three characteristic functions and of their second derivatives at \( s = 0 \):

\[
\begin{align*}
\hat{\lambda}_m^{(0)}(h) &= \sum_{h=1}^{n_m} W_{m,h} \lambda_0(0,h) \\
&+ \sum_{h=n_m+1}^{n} [A_0(h)h^3 + B_0(h)h^3] \lambda_0(0,h)
\end{align*}
\]

\( m = 1 \ldots n_m \) (II-5-13)

with (see II-3-11b):

\[
\begin{align*}
\lambda_0(0,h) &= [\Phi(s,h)/s]_{s=0} = 2(-1)^h - 1/(h) \\
\lambda_1(0,h) &= [\Phi(s,h)/s]_{s=0} \\
&= 2(-1)^h(h^2 - 6)[3(h)ds]^3.
\end{align*}
\]

(II-5-14)  

(II-5-15)

The general expression for such quantities which are linear combinations of \( \{X_j\} \) can be written:

\[ R = \Lambda \cdot X. \]  

(II-5-16)

The expression of the uncertainty in \( R \) is a generalization of (II-5-11) (see Anderson, 1958, pp. 19-20 & 25):

\[
\sigma^2_{\text{indep}, R} - Q\cdot \Lambda - Q = Q\cdot \Lambda^o - Q. 
\]

(II-5-17)

We have avoided so far making assumptions about the function \( t(X) \), apart from supposing that it is much flatter than \( q(Y_0|X) \) (see above). In some cases, however [see examples in Ramakrishnan & Moore (1982) and in § III. 1], it can be useful to introduce some additional mathematical linear constraints equivalent to a restriction on the distribution \( t(X) \). From a formal viewpoint, these restrictions can be considered as independent multiplicative terms in the expression for \( t(X) \):

\[
\begin{align*}
t(X) &= t_1(X)c(X) \\
c(X) &= \delta(C \cdot X - C_0)
\end{align*}
\]

(II-5-18)  

(II-5-19)

If \( t_1(X) \) is still 'flat' (in the sense discussed above) then the problem reduces to searching for the minimum and evaluating the uncertainty of the function,

\[
q_1(X) = q(Y_0|X)c(X)
\]

(II-5-20)

which is then equivalent (see Radhakrishna Rao, 1973, pp. 232–233) to solving the system,

\[
A_1X_0 = -B_1
\]

(II-5-21)

where a dummy unknown term \( X_{J+1} \) is added to the \( J \) components of \( X \), and where:

\[
A_1 = \begin{pmatrix} A & C \\ C & 0 \end{pmatrix}; \quad B_1 = \begin{pmatrix} B \\ C_0 \end{pmatrix}
\]

(II-5-22)

This transformation is recursive, so that several constraints will simply result in \( C \) becoming a matrix, and \( C_0 \) a vector.

When the simple linearity hypothesis is not satisfied, then the optimum \( X_0 \) under constraint can be searched for by the same kind of iterative algorithm as (II-5-5), with the difference that step (b) involves the resolution of (II-5-21) instead of (II-5-4). The uncertainty in a quantity \( R \) which is, at least locally, a linear combination of the \( \{X_j\} \) is then the same as (II-5-17), with \( \Lambda_0 \) and \( Q \) replaced respectively by (see Radhakrishna Rao, 1973, pp. 232–233)

\[
\begin{pmatrix} A_o & C \\ C & 0 \end{pmatrix}; \quad Q_1 = \begin{pmatrix} Q \\ 0 \end{pmatrix}
\]

(II-5-23)

In the remainder of this work, we shall use the word ‘uncertainty’ to refer to the \( \sigma_{\text{indep}} \) relative to each computed quantity, by using the square root of the second member in (II-5-17) with the above possible substitution.

III. An illustration: a low density serum lipoprotein (LDL)

III.1. Data analysis

By way of illustration we applied the mathematical treatment described above to the X-ray scattering data obtained with a low density serum lipoprotein extracted from hyper-lipidemic Rhesus monkeys (Luzzati, Tardieu & Aggerbeck, 1979). The X-ray scattering study was performed in water containing variable amounts of NaBr (up to 6M), over a solvent electron-density range 0.335 < \( \rho < 0.446 \) e/\( \AA^3 \). The X-ray source was a rotating-anode tube, the detector a linear position-sensitive proportional counter. The scattering experiments were performed at two settings: in one the sample-detector distance was 950 mm [14 experiments \( (900 \ \AA)^{-1} < s < (68 \ \AA)^{-1} \)], in the other 350 mm [10 experiments \( (190 \ \AA)^{-1} < s < (20 \ \AA)^{-1} \)]. All precautions were taken to ascertain the validity of the invariant-volume hypotheses (Luzzati et al., 1979; Luzzati & Tardieu, 1980).

To the best of our knowledge, this is one of the most complete and accurate solution X-ray scattering studies performed so far. Yet the experiments were carried out in 1976, at a time when our computing facilities were still rather primitive; as a consequence, the raw data were no longer in store, and the intensities available had suffered from a series of manipulations: background and solvent-scattering subtraction, smoothing, correction for collimation distortions. This circumstance is unfortunate since the algorithms derived above display their full force when applied to raw data, unspoiled by ill-controlled manipulations. On the other hand we find it appropriate to illustrate a new algorithm with data which have previously been analyzed using more conventional techniques.

One of the problems is to estimate the variance of the intensity data. If \( Y_{\mu,\sigma} \) is the 'observed' value at the \( kth \) position in the \( \mu \)th sub-experiment performed at solvent density \( \rho_{\mu,\sigma} \), after background and solvent-
scattering subtraction, smoothing and correction for collimation distortions, its variance \( \sigma_{Y_{\lambda,\alpha}}^2 \) can be roughly approximated (see discussion in § II-2) as the product of the 'count' \( Y_{\lambda,\alpha} \) by a smooth corrective function \( G(\hat{\rho},s) \) which takes into account the effects of the various manipulations. We determined this function by an iterative procedure, based upon the minimization of the local value of \( S_e^2 \) (see II-5-6) in the space \((\hat{\rho},s)\). By the virtue of this operation the overall value of \( \chi^2_e \) (II-5-8) dropped by several orders of magnitude, to a final value of 2.5.

At this stage, when we performed algorithm (II-5-5) we were surprised to observe in the autocorrelation function \( p(r,\hat{\rho}) \) extended negative regions, at large \( r \) and high \( \hat{\rho} \); this was a consequence of \( p_1(r) \) being negative (see Fig. 1). Although one could explain the negative regions of \( p(r,\hat{\rho}) \) by invoking interactions between LDL and \( Br^- \) [formally equivalent to \( v_1(r) \) taking negative values (Luzzati & Tardieu, 1980)] a simpler and more likely explanation can be sought in the presence of residual correlation effects, as indeed the observed intensities were not properly extrapolated to vanishing concentration. In the case of quasi-spherical objects like LDL the correlation effects can be tentatively described, at least at an ionic strength sufficiently high to screen electrostatic interactions, by the model of impenetrable spheres (Guinier & Fournet, 1955; Pilz, Glatter & Kratky, 1979):

\[
i_\alpha(s,\hat{\rho}) = c_e \left[ \frac{i_\alpha(s,\hat{\rho})}{c_e} \right]_{x \to 0} \times \left[ 1 - 24c_e \frac{\sin(4\pi Rs) - 4\pi Rs \cos(4\pi Rs)}{(4\pi Rs)^3} \right].
\]

(III-1-1)

where \( c_e \) is the volume concentration and \( R \) the radius of the spheres. Adopting \( R = 108 \) Å (see Luzzati et al., 1979), the correction factor (rightmost bracket in III-1-1) turns out to be quite small; yet its effect on \( p(r,\hat{\rho}) \) and consequently on \( p(r,\hat{\rho}) \) is to remove the negative regions (see Fig. 1). We thus adopted this correction in all subsequent calculations.

The properties of the scattering and autocorrelation functions at and very near the origins of \( s \) and \( r \) are simply related to specific structural parameters. Making the assumptions that the particles are homogeneous in density, that the position of the centre of mass is independent of the density of the solvent, that the invariant volume hypothesis is fulfilled and that the function \( v_1(r) \) is equal either to 1 or to 0, then these properties are expressed by equations (34) to (44) in Luzzati & Tardieu (1980). If the assumptions above are not fulfilled, then additional terms appear in these equations (see Sardet, Tardieu & Luzzati, 1976; Luzzati & Tardieu, 1980); yet in no case can \( i_\mu(0) \) be negative, since \( i_\mu(0) \) is a virtual experimental intensity.

The observed value of \( i_\mu(0) \), in fact, turns out to be slightly negative (approximately 4.5 times the uncertainty, see II-5-17), a paradox which can be explained by the presence, at low ionic strength, of residual electrostatic interactions. Whatever its cause, we decided to eliminate this anomaly by introducing the constraint \( i_\mu(0) = 0 \) (II-5-13, II-5-14) into (II-5-2) and (II-5-22). We also introduced the constraint \( i'_\rho(0) = 0 \) (II-5-13, II-5-15) although the small observed value for \( i'_\rho(0) \) was not anomalous. To determine the value of \( \rho_1 \) we used the condition

\[
i_{\mu_1}(0) = 0,
\]

(III-1-2)

which can be fulfilled by changing the value of \( \rho_1 \) and reconditioning the whole sets \( \{X_{\mu,\hat{\rho}}\} \) and \( \{X_{\mu,\hat{\rho}}^{\prime}\} \).

III.2. Determination of \( D_m \) and final results

The mathematical treatment requires the autocorrelation function to vanish beyond \( D_m \) and the parameter \( D = 1 \) (2\( As \)) to be equal to or larger than \( D_m \). Yet, from the standpoint of information theory 1 (2\( As \)) = \( D_m \) is a singular point, since, for that value and only for that value, the \( \{X\} \) parameters fulfill the requirements of being "degrees of freedom", i.e. to be the minimum number of independent parameters necessary and sufficient to describe the entire scattering curve (see § II.2). Therefore, the maximal chord of the particle, \( D_m \), is an important parameter both from the physical and the informational standpoints (Luzzati et

![Fig. 1. Correction of the residual correlation effect (see § III.1). The figure shows one portion of the function \( rp(r) \) m, determined with \( D_m = 305 \) Å. Dotted line: curve obtained before the correction of residual correlation effects; full line: curve obtained after that correction. Note that after the correction all negative values are removed and that, for \( r > 300 \) Å, all values are almost within the error bar [equal to the uncertainty as defined in (II-5-17)]. It is worth mentioning that the functions \( p_{\mu}(r) \) and \( p_{\rho}(r) \) are barely sensitive to this correction (results not shown).](image-url)
In order to determine $D_m$ one can perform algorithm (II-5-5) for different values of $D$ and then search the smallest $D$ for which the solution is consistent with the physical and the mathematical properties of the system (see also Moore, 1980). We scanned a range of $D$ which we expected to enclose $D_m$ (see Luzzati et al., 1979), and we performed algorithm (II-5-5) within the framework (parameters, variance estimations, additional constraints) specified above; we also assumed $\{B_i\} = 0$. For each value of $D$ the mathematical treatment provided the following information:

(a) The tests of the agreement between data and mathematical model: the reduced average square deviation $S^2$ (see II-5-6) and the reduced $\chi^2$ (see II-5-8).

(b) The most probable values of the scaling factors $\{K_{i,p}\}$.

c) The buoyant electron density $\tilde{\rho}_i$ (see III-1-2).

d) The most probable values of the $\{X_{i,k}\}$ and of the $\{A_j\}$, with their associated uncertainties.

e) Some quantities which are linear functions of $X$, namely: $i_r(0)/m, i_r'(0)/m, \tilde{i}_r(0)/m$ (II-5-13, II-5-14, II-5-15), $p_r(r)/m, p_r(r)/m, p_r(r)/m$ (II-5-12).

Let us examine first the plots of $S^2$ vs $D$ (see Moore, 1980), and $\chi^2$ vs $D$. As $D$ decreases, we expect these curves to display first a flat region, and then to rise sharply at $D = D_m$. This trend is indeed observed (Fig. 7) although the breaking point is not as sharp as one would desire. The safe conclusion to draw from this figure is that $D_m$ is larger than 280 Å.

We can now look at the autocorrelation functions. What we expect to observe is that any function $p(r,D)$ calculated with $D > D_m$ vanishes in the range $D_m < r < D$ and coincides with $p(r,D_m)$ when $r < D_m$. We plotted in Fig. 3 the outer region of three autocorrelation functions; each curve corresponds to one value of $D$. As expected, all the curves for $D$ larger than approximately 305 Å merge into (or barely exceed) the uncertainty over the range 305 Å $< r < D$. Therefore, the properties of the autocorrelation functions are consistent with $D_m$ being approximately equal to 305 Å. Moreover the presence in many curves $p(r,D)$ (not shown) and $i_r(0,D), i_r'(0,D)$, $\tilde{i}_r(0,D)$ (Fig. 4) of a maximum or a minimum as functions of $D$ in the vicinity of 305 Å indicates that the results obtained with $D = 305$ Å are the least sensitive to the choice of $D$. For all these reasons we feel it justified to adopt $D_m = 305$ Å. This value, however, is not based upon a straightforward determination and its accuracy is not easy to assess in unambiguous statistical terms.

Although our purpose in this work is not to discuss the structure of LDL - a more fruitful analysis must await a reliable set of raw data, properly extrapolated to vanishing concentration - it is worth noting that the results obtained with the new algorithm are in agreement with those obtained previously and that the main conclusions regarding the structure remain unchanged (Luzzati et al., 1979). The most significant novelty is the trustworthy definition of the uncertainties of the various parameters and of the autocorrelation functions; the correction of the autocorrelation effects, the introduction of the constraint $i_r(0) = 0$, the determination of $D_m$, are all based upon the estimate of statistical significance.

IV. Discussion

As we pointed out in the Introduction, we deal in this paper with only three aspects of the problem of information content and retrieval: the choice of the degrees of freedom, the process of data reduction and the estimation of the probability distribution of the results. Several previous papers have dealt with some of these problems: Damaschun et al. (1968), Luzzati et al. (1979), Luzzati (1980), Moore (1980), Müller et al. (1980) [we can also mention Glatter (1977) although his approach involves the use of spline functions, an approximation whose mathematical consequences should be properly analysed]. We strove in this paper to deal with the statistical problems with all possible care, and thus lay the ground for future work. Besides, we can point out a few aspects of our work - for example the definition of information content in scattering studies, the processing of entire complex data series in a single operation, the analysis of the determination of $D_m$ - which have not been discussed previously.

The parametrization we advocate in this work incorporates in a single algorithm all of the experi-
mental observations, recorded in different experimental conditions (e.g., camera settings, solvent densities, concentrations, etc.) including subsidiary measurements like those involved in the evaluation of normalization factors (a similar problem has been analysed recently by Ramakrishnan & Moore, 1982) and results in the storage of all that information in a finite number (II-3-2, II-3-3) of independent parameters, associated with a probability distribution matrix. This parametrization is based upon three hypotheses:

(a) The autocorrelation function has a finite support.

(b) The asymptotic trend of the intensity curves is known.

(c) The invariant volume hypothesis is fulfilled.

The hypothesis that the autocorrelation function exhibits a finite support is crucial: hardly any parametrization is possible unless some limit is set to the dimensions of the solute molecules (with the exception of a few special cases, for example rod-like or lamellar particles). The parametrization of the asymptotic trend that we adopt in this work (see II-3-2) is based upon semi-empirical considerations; other types of parametrization can be used, provided a regular behaviour of the intensity curves is ensured beyond \( s_m \) (it is also possible to allow for fluctuations around an average trend). The invariant-volume hypothesis can be regarded as an empirical parametrization of the solvent density dependence, which must be tested and used within the experimental range of solvent densities. Any extrapolation to lower or higher solvent densities involves physical models (Luzzati & Tardieu, 1980).

The process of data reduction is based upon a least-squares algorithm whose correct application deserves a few comments. With regard to the variance of the observations we point out in § II.1 that as a rule \( \sigma_Y^2 \) is a function of \( X \); in counting experiments for example \( \sigma_Y^2 \) is equal to the expected count \( Y_e(X) \) and not the observed count \( Y_o(X) \). One excuse for disregarding the \( X \) dependence is to avoid upsetting the linear form of the equations; a further justification resides in the fact that the mathematical approximations involved are of the second order and become negligible when the relative errors are small. More serious errors come from the various manipulations which the raw data may undergo before the statistical analysis: subtraction of blank scattering, correction for instrumental distortions (mainly collimation and polychromatism), smoothing (namely filtering of the statistical noise), extrapolations and interpolations. These manipulations introduce correlations between the differences \( \{ Y_o - Y_e(X) \} \) which are assumed to be statistically independent in the data and error analyses (see II-1-4 and II-5-6). Therefore, it must be emphasized that the only sound procedure is to apply the statistical treatment to the raw observations and to compensate the various distortions by means of additional terms in the calcu-

---

Fig. 3. Three-dimensional plot of the functions \( r \sigma_p(r) m \) relevant to the three characteristic functions (the vertical bars represent the uncertainty), at \( r > 200 \text{ Å} \). These curves are discussed in § III.2.
lated expression for $Y_{ij}(X)$. Besides, other experimental quantities known with finite accuracy (for example the scaling factors in § II.4) can also be parametrized and the new parameters treated as additional unknowns (see also Ramakrishnan & Moore, 1982).

It is also worth stressing that as a rule the statistical analysis involves both the experimental observations and the a priori knowledge one has of $X$. A determination of the most probable values of the degrees of freedom (and of the associated uncertainties) based upon the experimental observations alone is justified only when the experiment can be assumed to be much more accurate than the a priori knowledge, cannot be trusted. It is also rewarding to note that, in spite of all the manipulations, the LDL data still contain clear indications of residual correlation effects which had passed unnoticed in the previous analysis, short of a reliable determination of the statistical uncertainties. With regard to the determination of $D_{m}$, LDL probably is an unfavourable case since, as a consequence of the deeply corrugated surface (Luzzati et al., 1979) the autocorrelation function exhibits a smooth edge. Nevertheless, even in more favourable situations, it is unlikely that the determination of $D_{m}$ will become a plain straightforward operation.

We have laid the bases here for an analysis of the information content: this will be the subject of a forthcoming paper.

The authors are indebted to Drs Jean Courseul, Xavier Guyon, Bernard Prum, Françoise Schaeffer and Professor Gabriel Ruget for helpful discussions.

This work was partly supported by a grant from the Délégation Générale à la Recherche Scientifique et Technique.

References


