Acta Cryst. (1986). A42, 81-83

Fourier *versus* Hermite Representations of Probability Distributions

BY A. J. C. WILSON

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

(*Received* 24 *July* 1985; *accepted* 27 *September* 1985)

Abstract

Finite Fourier approximations fitted by use of orthogonality properties are identical with Fourier approximations fitted by least squares, but finite Gram-Charlier or Edgeworth (Hermite) approximations fitted by the two methods differ. This may give a partial explanation of the better performance of Fourier expressions for the probability distributions of structure factors. An alternative Hermite (Myller-Lebedeff) expansion exists, for which the orthogonality and the least-squares coefficients are identical, but they are not readily evaluated.

Introduction

Non-ideal probability distributions of structure factors have for long been represented by expansions in Gram-Charlier or Edgeworth series involving Hermite polynomials (Karle & Hauptman, 1953; Rogers & Wilson, 1953; for other references see Shmueli & Wilson, 1981), More recently, they have been expanded in Fourier or Fourier-Bessel series (Weiss & Kiefer, 1983; Shmueli, Weiss, Kiefer & Wilson, 1984; Weiss, Shmueli, Kiefer & Wilson, 1985). In general, the Fourier representations have been found to be better than the Hermite. At first sight this is rather surprising. The central-limit theorem applied to structure factors leads to the normal distribution, which forms the first term of the Gram-Charlier or Edgeworth series. Many observed distributions are nearly normal, and they have no obvious resemblance to a Fourier series. There is a trivial reason for the better performance of the Fourier representation; one can include as many Fourier terms as one desires, whereas only about five terms are readily available for the Hermite representation. There is, however, what may be a more fundamental reason.

Determination of coefficients

Consider the approximation of a function by a series of other functions:

$$f(x) \simeq \sum_{j=0}^{J} a_j g_j(x), \qquad (1)$$

where the number of terms is not infinite - the series is truncated at some maximum value J of j, and possibly 'censored' by the omission of some terms with $j<J$. In the present context, $g_j(x)$ could be $\cos 2\pi jx$ or $n(x)He_j(x)$, where $n(x)$ is the normal (Gaussian) distribution and $He_j(x)$ is a Hermite polynomial. The coefficients a_j could be evaluated in two ways:

Method A. They could be found by using orthogonality conditions; for example, if (1) is multiplied by a function $h_k(x)$ orthogonal to $g_j(x)$ for $j \neq k$ and integrated over the range of existence of $f(x)$, one has

$$\int f(x)h_k(x)\,dx \simeq \sum_{j\neq k} 0 + a_k \int g_k(x)h_k(x)\,dx, \qquad (2)$$

thus determining a_k.

Method B. Alternatively, the values of a_j could be chosen so that the sum in (1) is the best least-squares representation of $f(x)$:

$$\int [f(x) - \sum_j a_j g_j(x)]^2\,dx \qquad (3)$$

should be a minimum with respect to the a_j. This leads to a set of simultaneous equations for determining the a_j:

$$\int g_k(x)[f(x) - \sum_j a_j g_j(x)\,dx = 0. \qquad (4)$$

For the Fourier series the results of the two methods are *identical*, since the cosines are orthogonal to each other; the simultaneous equations reduce to a set of individual equations each determining one coefficient. The same is true for any set of functions for which $h_k(x)$ is the same as $g_k(x)$ (see, for example, Spiegel, 1974, Ch. 3). For the Hermite series (whether Gram-Charlier or Edgeworth) $h_k(x) \neq g_k(x)$:

$$g_k(x) = n(x)He_k(x) \qquad (5)$$

but

$$h_k(x) = He_k(x). \qquad (6)$$

The coefficients a_j found by method A are (comparatively) easily evaluated in terms of simple functions of the moments of $f(x)$, and of even simpler functions of its cumulants, but method B leads to a set of simultaneous equations for the a_j, and the a_j have no simple physical interpretation.

0108-7673/86/020081-03$01.50

© 1986 International Union of Crystallography

Example

These points may be illustrated by the simplest possible Gram–Chalier or Edgeworth (Hermite) approximation:

$$f(x) \simeq n(x)[1 + aHe_4(x)], \tag{7}$$

applicable for a symmetrical function such as the probability distribution of the structure factors of a centrosymmetric structure. Method A leads immediately to

$$a = \int f(x)He_4(x)\,dx / \int n(x)He_4^2(x)\,dx \tag{8}$$

$$= k_4/4!, \tag{9}$$

where k_4 is the fourth cumulant of $f(x)$. Method B leads to

$$a = \int f(x)n(x)He_4(x)\,dx / \int n^2(x)He_4^2(x)\,dx, \tag{10}$$

which has an extra factor of $n(x)$ in both numerator and denominator, reducing both in comparison with those of (8). The reduction is drastic; the integral in the denominator of (8) has the value $4! = 24$, but that in (10) has the value $105/32\pi^{1/2} = 1{\cdot}85\ldots$. It is not obvious whether method A or method B will give the larger value of a, but it is obvious that they will give *different* values. It is also obvious that a from method B has no simple relationship to the moments of $f(x)$, and indeed that it has no simple physical interpretation.

Least-squares interpretation of the orthogonality fit

Equation (1), with $g_j(x) = n(x)He_j(x)$ and a_j evaluated by using the orthogonality relations, is not a least-squares fit to $f(x)$. Two questions immediately present themselves:

(1) Is (1) a least-squares fit to anything, and if so, to what?

(2) Is there an expansion, with $n(x)$ as its leading term, that gives a least-squares fit to $f(x)$?

Contemplation of variations on (3) gives a positive answer to both questions.

One sees readily that minimizing

$$\int [f(x)n^{-1/2}(x) - n^{1/2}(x)\sum a_jHe_j(x)]^2\,dx \tag{11}$$

with respect to the a's leads to the same set of equations for the a's as does method A. The usual Gram–Charlier/Edgeworth expansions thus correspond to a least-squares fit to the function

$$n^{-1/2}(x)f(x), \tag{12}$$

which decreases much more slowly with x than does $f(x)$. In the crystallographic application, therefore, method A gives great weight to fitting the distribution of the (comparatively few) strong reflexions, at the expense of a poorer fit for the distribution of the (much more numerous) weaker ones. This is readily seen by rewriting (11) in the form

$$\int n^{-1}(x)[f(x) - n(x)\sum a_jHe_j(x)]^2\,dx. \tag{13}$$

The mean-square 'distance' between the distribution $f(x)$ and its representation is thus magnified by the reciprocal of the normal distribution, which becomes very large for x large.

Cramér (1945) gives a criterion for the existence of a convergent Gram–Charlier series. Put crudely, $f(x)$ must go to zero so fast that the integral of $f(x)n^{-1/2}(x)$ is finite. Expression (12) gives a physical picture of this criterion; if it is not satisfied the function to be fitted by least squares increases indefinitely with increasing x, and thus cannot be matched by a series of functions that ultimately go to zero.

To return to the second question, if in (1) $He_j(x)$ is replaced by $He_j(2^{1/2}x)$, so that the representation is

$$f(x) \simeq \sum a_jn(x)He_j(2^{1/2}x), \tag{14}$$

the orthogonality condition is satisfied with

$$h_k(x) = g_k(x) = n(x)He_k(2^{1/2}x), \tag{15}$$

and the values of a_k given by methods A and B become identical, just as for the Fourier representation. Both give

$$a_k = \int f(x)n(x)He_k(2^{1/2}x\,dx / \int n^2(x)He_k^2(2^{1/2}x)\,dx \tag{16}$$

$$= (2\pi^{1/2}/n!)\int f(x)n(x)He_k(2^{1/2}x)\,dx. \tag{17}$$

One sees readily that $He_k(2^{1/2}x)$ corresponds to the alternatively defined Hermite polynomial $H_k(x)$, from which it differs only by a factor of $2^{n/2}$ (Abramowitz & Stegun, 1964, formulae 26.2.31–32). As in the Fourier representation, the coefficients a_k are not readily expressible in terms of the moments, but (14) would give a least-squares fit to $f(x)$ without the undesirable overweighting of the values for large x. Preliminary attempts to evaluate a_k by integration of (17) with $f(x)$ in the exact form given by Kluyver (1907) have not been encouraging.

Expansions of the type (14) were considered by Myller-Lebedeff (1907). Her work appears to have been almost forgotten by statisticians; it rates a three-line footnote in Kendall & Stuart (1977, p. 90). Titchmarsh (1937, p. 79) gives a paragraph to the expansion, but without reference to its origin or uses. Perhaps the most familiar application is in the wave-mechanical representation of a simple harmonic oscillator (see, for example, Prince, 1982).

References

ABRAMOWITZ, M. & STEGUN, I. A. (1964). *Handbook of Mathematical Function.* Washington, DC: US Government Printing Office.

CRAMÉR, H. (1945). *Mathematical Theory of Statistics.* Uppsala: Almqvist & Wiksells.

KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131-135.
KENDALL, M. G. & STUART, A. (1977). *The Advanced Theory of Statistics*, Vol. 1, 4th ed. London: Griffin.
KLUYVER, J. C. (1907). *Proc. K. Ned. Akad. Wet.* **8**, 341-350.
MYLLER-LEBEDEFF, W. (1907). *Math. Ann.* **64**, 388-416.
PRINCE, E. (1982). *Mathematical Techniques in Crystallography and Materials Science.* New York: Springer.
ROGERS, D. & WILSON, A. J. C. (1953). *Acta Cryst.* **6**, 439-449.
SHMUELI, U., WEISS, G. H., KIEFER, J. E. & WILSON, A. J. C. (1984). *Acta Cryst.* A **40**, 651-660.
SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* A **37**, 342-353.
SPIEGEL, M. R. (1974). *Fourier Analysis with Application to Boundary Value Problems. Schaum's Outline Series.* New York: McGraw-Hill.
TITCHMARSH, E. C. (1937). *Introduction to the Theory of Fourier Integrals.* Oxford: Clarendon Press.
WEISS, G. H. & KIEFER, J. E. (1983). *J. Phys. A*, **16**, 489-495.
WEISS, G. H., SHMUELI, U., KIEFER, J. E. & WILSON, A. J. C. (1985). In *Structure and Statistics in Crystallography*, edited by A. J. C. WILSON, pp. 23-42. Guilderland, NY: Adenine Press.

Acta Cryst. (1986). A**42**, 83-85

Molecular Speleology: The Exploration of Crevices in Proteins for Prediction of Binding Sites, Design of Drugs and Analysis of Surface Recognition

BY ARTHUR M. LESK*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

(*Received* 24 *June* 1985; *accepted* 8 *August* 1985)

Abstract

A method is described for analyzing molecular-surface complementarity, including the binding of ligands to proteins or the interaction of elements of secondary structure in protein interiors. A computer program can identify and model molecules that satisfy general criteria for good binding affinity. Computational tests are presented. This approach is likely to have useful application in the analysis of surface recognition in proteins, including the identification of binding sites, and in the design of drugs for specific targets, by (i) suggesting potential pharmacophores to the medicinal chemist for further computational analysis or laboratory testing, (ii) suggestion of derivatives of a known ligand to enhance its affinity, or (iii) searching a data base of known drugs for a match to the predicted ligand.

Introduction

Emil Fischer first proposed the 'lock and key' model of enzyme-substrate interactions. We now recognize the importance of surface complementarity not only for ligand binding, but for the interactions of packed α-helices and β-sheets in protein interiors which are crucial in stabilizing native conformations (Lesk, 1981; Chothia, 1984). Important applications of computational methods for analyzing molecular complementarity include:

(1) Analysis of the packing in protein interiors: What will be the effect of a mutation on the conformation of a protein (Lesk & Chothia, 1980)? What freedom do packed secondary structures have to facilitate and transmit conformational changes (Chothia, Lesk, Dodson & Hodgkin, 1983)?

(2) Prediction of ligands complementary to specific clefts in proteins. Can we thereby design drugs of high affinity and specificity (Tickle, Sibanda, Pearl, Hemmings & Blundell, 1984; Beddell, 1984)? Can we rationalize the specificities of antibodies? With the application of protein-engineering techniques to antibodies, it will be useful to analyze changes in the antigen-binding site (Neuberger, 1983).

Given a protein structure that contains a cleft, how can one identify a ligand that has a structure complementary to the cleft? Analyses of protein-ligand interactions suggest that loss of solvent-accessible surface area, and complementarity in shape and charge distribution are the major determinants of affinity and specificity (Janin & Chothia, 1978; Chothia, 1984; Kollman, 1984). Studies of complementarity have used physical models [including making casts, using known protein structures as molds (Blow & Smith, 1975)], empirical parameters characterizing hydrophobicity (Smith, Hansch, Kim, Omiya, Fukumura, Selassie, Jow, Blaney & Langridge, 1982) and interactive computer graphics (Langridge, Ferrin, Kuntz & Connolly, 1981; Busetta, Tickle & Blundell, 1983).

We describe here a computational technique to explore clefts in proteins and suggest candidate ligands. It does not require the facilities of interactive-graphics packages, but could easily and profitably be integrated with them. [This problem should be distinguished from a related one: determining the

* Permanent address: Fairleigh Dickinson University, Teaneck-Hackensack Campus, 1000 River Road, Teaneck, NJ 07666, USA.

0108-7673/86/020083-03$01.50 © 1986 International Union of Crystallography