Submitted to the Statistical Science

Bayesian Treatment of Negative Intensity Measurements in Crystallography*

Simon French and Keith S. Wilson

University of Warwick and University of York

Abstract. The true intensities of reflections, i.e. diffracted rays, in X-ray crystallography are known to be non-negative. However, the experimental value of the intensity, being measured by a difference between background and peak, can sometimes be negative for small reflections. This is particularly true for large biological molecules where many intensities may be small. Until the mid 1970s, non-Bayesian methods effectively set negative measurements to zero, introducing a positive bias into the resulting intensity distribution and perturbing the electron density maps and refined molecular models. In terms of an application of Bayes Theorem the problem was simple: the use of a suitable prior distribution would enforce non-negativity of the estimated intensities. Computationally, the problem was not quite so trivial, but with suitable approximation and tabulation Bayesian estimates were produced. It was instrumental in introducing Bayesian ideas into a 'hard physical' science. For small molecules a theoretical justification of the prior that we used had been known since 1949. However, we had to conjecture that for large biological molecules such as proteins a similar distribution form for the prior would be appropriate. Empirically our results justified our conjecture. Moreover, subsequently it was shown using central limit theorems for correlated variables that our conjecture was justified. The method rapidly became widely used within crystallographic data analysis, the "truncation" algorithm being widely applied in macromolecular crystallography packages over the last 25 years and our original paper is still highly cited.

Key words and phrases: Bayesian methodology, crystallography, Wilson statistics.

1. INTRODUCTION

This Bayesian application dates back to the late 1970s, but its longevity as a standard method of crystallography shows its strength. Moreover, the application

Department of Statistics, Coventry, CV4 7AL, UK simon.french@warwick.ac.uk
Department of Chemistry, Heslington, York, YO10 5DD, UK
keith.wilson@york.ac.uk

*Originally our work was supported by the Hayward Foundation (SF) and ICI (KW).

played a major role in the acceptance of Bayesian methodology within mainstream crystallographic statistics within a decade of its publication.

X-ray crystallography is long established as the prime technique for determining a molecule's 3D structure, nowadays recognized as an essential part of the full understanding and description of molecules both large and small. Crystallography has underpinned the structural chemistry of the last century and of large biological molecular structures in the last 50 years. For a full introduction to crystallography, see Rupp, B. (2009).

Briefly, crystals are composed of repeating units, called *unit cells*, each containing one or a small number of molecules and arranged in a lattice. When X-rays are shone at a crystal, the lattice acts as a diffraction grating, scattering the beam into rays, known as *Bragg reflections*, radiating out in a fixed pattern: see Fig. 1. The intensities of these reflections are related to the moduli of the coefficients in the three-dimensional Fourier expansion of the electron density of molecule, and are thus the essential experimental data in the elucidation of 3D structures of molecules. Within crystallography the Fourier coefficients are known as *structure factors*.

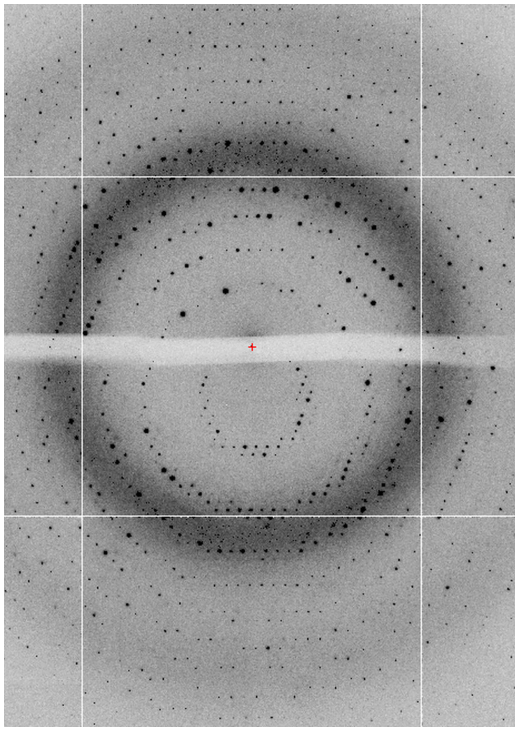

FIG 1. *A typical diffraction image from a protein. The image contains a regular array of Bragg reflections superimposed on a background of varying intensity. The integrated intensity of each peak represents the square of the amplitude of a Fourier component of the electron density in the crystal.*

Current systems for measuring X-ray intensities include imaging plates, charged coupled devices and pixel detectors. All estimate the intensity essentially by measuring the value in the reflection itself and then subtracting a background value near to it, Figure 1(b). Thus:

$$I = reflection\ strength - background \tag{1}$$

However, some intensities are small relative to the background and for large biological molecules the proportion of such reflections is often large, with the intensity/standard deviation ratio often dropping to close to 1.0 at the edge of the diffraction pattern. Thus it is inevitable that random errors lead to some intensities being measured as negative, notwithstanding that theory says they must be non-negative. At the time we developed our Bayesian approach in the late 1970s, most software packages simply set negative measurements to zero or, perhaps, some small positive number, giving rise to biased data sets. Instead of recognising that negative measurements provided strong information that the intensity was small, the issue was perceived as a 'problem' that got in the way of subsequent data analysis, in which the each intensity would need to be square-rooted to give the corresponding Fourier coefficient's modulus.

For a Bayesian, the 'problem' truly does not exist. That an intensity is non-negative is clearly and unarguably prior knowledge. So one simply chooses a prior which enforces non-negativity: i.e. one which is zero on the negative line.

2. THE BAYESIAN MODEL

A straightforward application of Bayes Theorem gives:

$$p_J(J|I) \propto_J p_I(I|J) \times p_J(J)$$

where $J = \|F\|^2$ is the 'true' intensity, which is the square of Fourier coefficient modulus; I is the measured intensity and the proportionality is as a function of J. Then in later calculation of the three-dimensional structure we may use for the intensity:

$$E_J(J|I) = \int_0^\infty J p_J(J|I) dJ$$
$$var_J(J|I) = \int_0^\infty [J - E_J(J|I)]^2 p_J(J|I) dJ$$

or for the modulus of the Fourier coefficient:

$$E_J(\|F\||I) = \int_0^\infty \sqrt{J} p_J(J|I) dJ$$
$$var_J(\|F\||I) = \int_0^\infty [\sqrt{J} - E_J(\|F\||I)]^2 p_J(J|I) dJ$$

The measured intensity I is formed as in (1), but with further operations applied, including corrections for Lorentz, polarization, absorption, extinction and radiation-damage effects. Moreover, the crystal symmetry may imply that certain Fourier coefficients must have the same moduli, in which case the observed intensity will have been averaged over these 'equivalent reflections'. All corrections and averaging are assumed to have been carried out on the raw observations, be they positive or negative.

It is reasonable to assume that $p_I(I|J)$ is normal $N(J, \sigma^2)$, where σ^2 is known. The raw observations are formed from a difference of Poisson distributed intensities, and the averaging and corrections drive the measured intensities towards normality to a good approximation. With careful experimental practices, unbiasedness is a fair assumption and empirical estimates of σ^2 are good.

One obvious choice for the prior is

$$p_J(J) = \begin{cases} 1 & \text{if } J \geq 0 \\ 0 & \text{if } J < 0 \end{cases}$$

However, there is more information available. Wilson (1949)[1] showed that, in the case of *acentric* crystal symmetry:

$$p_J(J) = \begin{cases} \Sigma^{-1} \exp(-J/\Sigma) & \text{if } J \geq 0 \\ 0 & \text{if } J < 0 \end{cases} \tag{2}$$

while in the case of *centric* crystal symmetry:

$$p_J(J) = \begin{cases} (2\pi\Sigma J)^{-1} \exp(-J/2\Sigma) & \text{if } J \geq 0 \\ 0 & \text{if } J < 0 \end{cases} \tag{3}$$

In each case, Σ is the mean intensity over an appropriate subset or *shell* of reflections. In a strict sense, Σ is unknown, but the number of intensities in a shell is typically large. So taking an empirical Bayesian approach, we simply estimated it for each shell from the data.

Thus we have a Bayesian prescription for processing the data in a manner which ensures non-negative posterior means for the intensity and Fourier coefficient moduli. Computationally, particularly in the 1970s, there was still the issue of implementing the calculations, even though in many cases the posterior is effectively just a truncated normal distribution. We tabulated the integrals for $-4\sigma < I < 3\sigma$ and then interpolated in this for particular observations. Outside this range the posterior is effectively normal. The computations proved feasible then and most certainly are today. Details are in French and Wilson (1978).

French (1975) in his doctoral research developed a more sophisticated Bayesian analysis of the intensity profile at a reflection as a diffractometer stepped through the peak but this required considerable tailoring to the specifics of the diffractometer used: see also French and Oatley (1982); Oatley and French (1982). The simplicity of the approach in French and Wilson (1978) means that it can be applied to intensity data, be they collected by counters, photographs or other means. That, together with the quality of its results, has meant that it remains in use across crystallography to the present day.

3. THE IMPACT OF OUR METHOD

Scientifically, the greatest impact of our work is clearly that it is used. As we write this, the paper has over 600 citations and is still being cited. The vast majority of citations are from structural studies which have used the method in determining molecular structures. It is described in the International Tables for Crystallography (IUCr, 2001b, Section 7.5.6, page 661). The algorithm is implemented in several current crystallographic packages: e.g. TRUNCATE in CCP4 (http://www.ccp4.ac.uk/) and BAYEST in XTAL (http://xtal.sourceforge.net/).

Not all citations are from structural studies, though. In terms of Bayesian impact, it is pleasing to see that many relate to statistical methodology within

[1] A.J.C.Wilson and K.S. Wilson are not related

crystallography and, tracking back to the earlier of such citations, that our paper played a major role in acceptance of Bayesian methods. Remember that we did this work and published it in the 1970s. The Bayesian approach was seldom accepted then, and certainly not recognised in the physical sciences in which its explicit subjectivity was an anathema to the majority of researchers. Our paper was published in 1978. Ten years later, Schwarzenbach et al (1989) reported to the International Union of Crystallography on statistical methodology and Bayesian methods were presented alongside frequentist ones with equal status. Our paper was cited along with a more theoretical paper by French (1978) on the use of Bayesian hierarchical modelling as a framework for crystallographic refinement and the work of French and Oatley on the Bayesian method of intensity profile fitting (Oatley and French, 1982; French and Oatley, 1982). Within 10 years Bayesian methods were accepted, and it is clear that the simplicity and practical results of French and Wilson (1978) were central in achieving this acceptance. Other examples of the penetration of Bayesian thinking into crystallographic statistics and the role of French and Wilson (1978) in this are provided by Bricogne (1988) and Gilmore (1996). Moving up to date, (Rupp, B., 2009, Ch. 7) provides a modern introduction to crystallographic statistics in which not only do Bayesian statistics feature, but our method is used as a worked example for the student.

4. WILSON'S STATISTICS AND MACROMOLECULES

Equations (2) and (3) are known throughout crystallography as Wilson's Statistics. His original derivation was based on the assumptions that the atoms of a molecule were randomly and independently distributed within the unit cell, with respect to both (i) the distances between them and (ii) their relative orientation. These assumptions allowed him to use the central limit theorem to derive the probability distributions. For small molecules, the assumptions were reasonable; but for large biological molecules such as proteins, in which the atoms tend to be arranged in chains twisted into a ball or knot, they were very questionable. However, a lot of experience within protein crystallography had shown that while the behaviour of Σ over *different* shells of data was not as Wilson had originally predicted, though *within* a shell of data, Wilson's Statistics did apply. Our paper summarised the empirical empirical evidence of this (French and Wilson, 1978). We also postulated – and seem to have been the first to do so (IUCr, 2001a, Section 2.1.4.5, page 195.) – that Wilson's original 1949 analysis would generalise if he drew upon central limit theorems which dealt with correlated variables rather than the better known theorems for independent variables. Moreover, we postulated that the forms of his distributions would be unchanged, save for different values for Σ. We corresponded with Arthur Wilson on this and in 1981 he published a theoretical paper confirming our conjectures (Wilson, 1981).

Examining assumptions (i) and (ii) above more closely: (i) is clearly not valid as inter-atomic distances are governed by the nature of chemical bonds which have sets of preferred values. This departure from assumption is allowed by using empirically determined values for Σ for each shell of data. For most structures (ii) is a reasonable approximation to reality. Nevertheless in recent years it has become evident that there are structures for which (ii) breaks down, for example in crystals with systematic orientations of their contents through non-crystallographic

symmetry. Extensions to our approach are being developed by others to allow for such deviations, while following the basic assumptions of Bayesian statistics.

An area which remains to be addressed is the incorporation of Bayesian estimates of experimental error into statistical methods of phase determination. The phase information is lost during the recording of the structure factor amplitudes the so-called phase problem at the core of crystallography. The assumptions of positivity of the electron density and of atomicity impose restrictions on the phases given the amplitudes, sufficient for them to be determined using a set of statistical relations. However, the probabilities of these relations is dependent on the values of the amplitudes involved, and present methods do not take account of the experimental errors in these. A powerful extension of our Bayesian approach would combine the error estimates for the amplitudes into the statistical phase relations to provide improved posterior distributions and more meaningful indicators of when the phase problem is likely to be solvable given the data quality.

REFERENCES

Bricogne, G. and Wilson, K. (1988). A Bayesian Statistical Theory of the Phase Problem. I. A Multichannel Maximum-Entropy Formalism for Constructing Generalized Joint Probability Distributions of Structure Factors. *Acta Crystallographica* **A44** 517–545.

French, S. (1975). On the interpretation of X-ray crystallographic data on proteins. *D.Phil Thesis* University of Oxford.

French, S. (1978). A Bayesian three-stage model in crystallography. *Acta Crystallographica* **A34** 728-738.

French, S. and Oatley, S.J. (1982) Bayesian statistics: an overview. In S. Ramaseshan, M.F. Richardson and A.J.C. Wilson (Eds) *Crystallographic Statistics: Progress and Problems* Indian Academy of Sciences, 19–51

French, S. and Wilson, K. (1978). On the treatment of negative intensity observations. *Acta Crystallographica* **A34** 517–525.

Gilmore, C.J. (1996). Maximum Entropy and Bayesian Statistics in Crystallography: a Review of Practical Applications. *Acta Crystallographica* **A52** 561–589.

IUCr (2001a) *International Union of Crystallorgraphy: International Tables for Crystallography.* Volume B 2nd Edition. Kluwer Academic Publishers.

IUCr (2001b) *International Union of Crystallorgraphy: International Tables for Crystallography.* Volume C 2nd Edition. Kluwer Academic Publishers.

Oatley, S.J. and French, S. (1982). A profile-fitting method for the analysis of diffractometer data. *Acta Crystallographica* **A38** 537–549.

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology.* Garland Science, New York.

Schwarzenbach, D., Abrahams, S.C., Flack, H.D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R.E., Prince, E., Robertson, B.E, Rollett, J.S. and Wilson, A.J.C. (1989). Statistical Descriptors in Crystallography: Report of the International Union of Crystallography Subcommittee on Statistical Descriptors. *Acta Crystallographica* **A45** 63–75.

Wilson, A.J.C. (1949). The probability distribution of X-ray intensities. *Acta Crystallographica* **2** 318-321

Wilson, A.J.C. (1981). Can intensity statistics accommodate stereochemistry? *Acta Crystallographica* **A37** 808–810