AN IMPROVED PRIOR FOR IMAGE RECONSTRUCTION IN X-RAY FIBER DIFFRACTION

Shyamsunder Baskaran and R. P. Millane

Computational Science and Engineering Program
Purdue University, West Lafayette, Indiana 47907-1160
{shyam,rmillane}@purdue.edu

ABSTRACT

The structure completion problem in fiber diffraction is addressed from a Bayesian perspective. The experimental data are sums of the squares of the amplitudes of particular sets of Fourier coefficients of the electron density. In addition, a part of the electron density is known. The image reconstruction problem is to estimate the missing part of the electron density. A Bayesian approach is taken in which the prior model for the image is based on the fact that it consists of atoms, i. e. the unknown electron density consists of separated sharp peaks. The conventional prior assumes that the positions of the unknown atoms are uniformly distributed. We improve this prior by treating the positions of the known atoms as containing normally distributed coordinate errors. Currently used heuristic methods are shown to correspond to certain maximum *a posteriori* estimates of the Fourier coefficients. An analytical solution for the Bayesian minimum mean-square-error estimate is derived. Simulations show that the minimum mean-square-error estimate gives better results when the new prior is used.

1. INTRODUCTION

X-ray crystallography is used to study three-dimensional (3-D) molecular structures at atomic resolution [1]. The x-ray diffraction pattern from a crystalline (periodic) specimen consists of Nyquist spaced samples of the square of the Fourier transform magnitudes of one period of the image. The (3-D) support, $\mathcal{C}$, of one period of the image is called the unit cell. The data are therefore $|F_{\mathbf{h}}|^2$ where

$$F_{\mathbf{h}} = \int_{\mathcal{C}} \varrho(\mathbf{r}) \exp(i2\pi\mathbf{h}\cdot\mathbf{r})\, d\mathbf{r}, \qquad (1)$$

where $\varrho(\mathbf{r})$ is the electron density function and $\mathbf{h} \in \mathbb{Z}^3$. The $F_{\mathbf{h}}$ are known as structure factors, and may be expressed as $|F_{\mathbf{h}}|\exp(i\phi_{\mathbf{h}})$ in terms of the modulus and phase angle. Fiber diffraction studies are performed on substances that form aggregates of small crystallites that are randomly rotated about a preferred axis, resulting in cylindrical averaging of the diffraction pattern, so that the data are

$$I_j = \sum_{\mathbf{h}\in\mathcal{S}_j} |F_{\mathbf{h}}|^2, \qquad (2)$$

where the $\mathcal{S}_j$ are sets of points on the sampling lattice with the same cylindrical polar radius.

An important and practical problem that occurs in x-ray fiber diffraction involves completing the image function $\varrho(\mathbf{r})$ from the intensity data, and a partial image $\varrho^P(\mathbf{r})$ which may be obtained from a least-squares solution or from that of a previously solved structure of a similar molecule. This occurs in structural biology where the 3-D structure (image) consists of known (located in 3-D) components, and other unknown components

This work was supported by NSF (DBI-9722862).

0-8186-8821-1/98 $10.00 © 1998 IEEE

(such as other molecules, ions or solvent molecules) that need to be located [2, 3]. Denoting the missing contribution to the image by $\varrho^Q(\mathbf{r})$, we have that

$$\varrho(\mathbf{r}) = \varrho^P(\mathbf{r}) + \varrho^Q(\mathbf{r}), \tag{3}$$

so that

$$F_{\mathbf{h}} = F_{\mathbf{h}}^P + F_{\mathbf{h}}^Q, \tag{4}$$

and the problem reduces to one of estimating $\varrho(\mathbf{r})$, or equivalently $F_{\mathbf{h}}$, from the I_j and $F_{\mathbf{h}}^P$.

2. PRIOR MODEL AND DENSITY FUNCTIONS

From (2), we may deduce that the transformation from the data I_j to the image $\varrho(\mathbf{r})$, requires that we estimate the complex quantities denoted by the set $\{F_{\mathbf{h}}\}_j = \{F_{\mathbf{h}}, \mathbf{h} \in \mathcal{S}_j\}$ for each datum. The problem is underdetermined in general and is resolved by incorporation of prior information on the solution. The prior is based on a property of the electron density, *atomicity*, which allows for a representation in the form

$$\varrho(\mathbf{r}) = \sum_{j \in \mathcal{N}} \varrho_j(\mathbf{r} - \mathbf{r}_j), \tag{5}$$

where the $\varrho_j(\mathbf{r})$ are radially symmetric, positive, well behaved functions that correspond to the electron density of the jth atom positioned at the origin, $\mathbf{r}_j$ its position, and the set $\mathcal{N}$ indexes the atoms in the unit cell. This permits the structure factor to be represented as

$$F_{\mathbf{h}} = \sum_{j \in \mathcal{N}} f_j(\mathbf{h}) \exp\left(i\phi_j(\mathbf{h})\right), \tag{6}$$

where $f_j(\mathbf{h})$, known as the atomic scattering factor, is the Fourier Transform of $\varrho_j(\mathbf{x})$, and $\phi_j(\mathbf{h}) = 2\pi\mathbf{r}_j \cdot \mathbf{h}$. (The dependence on $\mathbf{h}$ is suppressed in the following.)

In the absence of other information, we assume statistical priors for the atomic positions of the sets of atoms of the known ($\mathcal{P}$) and missing ($\mathcal{Q}$) parts. The currently used prior in fiber diffraction [3] assumes that the positions of the atoms in the missing part of the structure, $\{\mathbf{r}_j, j \in \mathcal{Q}\}$, are uniformly distributed in the unit cell $\mathcal{C}$. The improved prior assumes in addition that the positions of the atoms in the known part $\{\mathbf{r}_j, j \in \mathcal{P}\}$ contain normally distributed coordinate errors.

$\{F_{\mathbf{h}}\}_j$ may be represented by a vector $\mathbf{Y}$, with $n_j = 2|\mathcal{S}_j|$ components corresponding to the real and imaginary parts of each constituent structure factor. Equation (2) becomes

$$I_j = \|\mathbf{Y}\|^2, \tag{7}$$

where the vector of structure factors is broken down as $\mathbf{Y} = \mathbf{\Theta} + \mathbf{T} + \mathbf{X}$, where $\mathbf{\Theta}$ is the contribution from the known part of the structure ($\{F_{\mathbf{h}}^P\}_j$), $\mathbf{T}$ is the contribution of the errors in the known structure and $\mathbf{X}$ is the contribution from the missing part of the structure ($\{F_{\mathbf{h}}^Q\}_j$). The components of $\mathbf{Y}$ are independent and identically distributed.

From the uniform density of $\{\mathbf{r}_j, j \in \mathcal{Q}\}$, the density function of a component of $\mathbf{X}$, X_i is $N(0, \Sigma_Q/2)$, where $N(a, b)$ denotes the normal pdf with mean a and variance b, and $\Sigma_Q = \sum_{j \in \mathcal{Q}} f_j^2$, is a parameter that is related to the amount of missing electron density. This represents the prior pdf for the missing part of the image.

The density function of $\mathbf{\Theta}$ is $\delta(\mathbf{\Theta})$, representing an exact known part. This prior is improved upon (following [4]), using a normal pdf $N(\mathbf{r}_j, C_{\mathbf{r}})$, where, $C_{\mathbf{r}}$ is a diagonal iid covariance matrix (with diagonal elements σ_P^2), for the coordinates $\{\mathbf{r}_j, j \in \mathcal{P}\}$. This results in the density function for a component of $\mathbf{T}$, T_i to be $N((D-1)\Theta_i, \Sigma_P(1 - D^2)/2)$, where Σ_P is defined similar to Σ_Q and $D = \langle\cos(2\pi\mathbf{h} \cdot \delta\mathbf{r}_j)\rangle$, is a parameter that quantifies the errors in the coordinates of the known partial structure. (Σ_P, Σ_Q and D are functions of $\mathbf{h}$.)

Since $Y_i = \Theta_i + T_i + X_i$, the pdf of Y_i is $N(D\Theta_i, \Sigma_U/2)$, where $\Sigma_U = \Sigma_Q + (1 - D^2)\Sigma_P$.

Assuming that there are no errors in the observations I_j, we obtain the posterior density function for $\mathbf{Y}$, given I_j as

$$P_{\mathbf{Y}|I_j}(\mathbf{y}) = (\pi\Sigma_U)^{-n_j/2}\exp(-\|\mathbf{y}-D\boldsymbol{\Theta}\|^2/\Sigma_U)\,\delta(I_j-\|\mathbf{y}\|^2). \quad (8)$$

3. ESTIMATORS

The posterior density function is the Bayesian solution to the inverse problem. The maximum entropy, or minimum mean-square-error (MMSE) estimate is the posterior mean, which gives [3]

$$\hat{F}_{\mathbf{h}}^{\mathrm{MMSE}} = \left(\frac{I_{n_j/2}(\chi)}{I_{n_j/2-1}(\chi)}\sqrt{\frac{I_j}{I_j^P}}\right)F_{\mathbf{h}}^P\,, \quad (9)$$

for $\mathbf{h}\in\mathcal{S}_j$, where $I_j^P=\|\boldsymbol{\Theta}\|^2$, $I_\nu(\cdot)$ is the modified Bessel function of the first kind of order ν and $\chi = 2D(I_jI_j^P)^{1/2}/\Sigma_U$. Note that implementing this estimate requires estimation of the parameters Σ_U and D.

Currently, two heuristic methods are used to solve this problem in x-ray fiber diffraction [2, 3]. We have shown that these methods correspond to two maximum *a posteriori* (MAP) estimates [3], denoted here by MAP1 and MAP2, which are given by

$$\hat{F}_{\mathbf{h}}^{\mathrm{MAP1}} = \left(\sqrt{\frac{I_j}{I_j^P}}\right)F_{\mathbf{h}}^P, \quad (10)$$

and

$$\hat{F}_{\mathbf{h}}^{\mathrm{MAP2}} = \left(\frac{1}{|F_{\mathbf{h}}^P|}\sqrt{\frac{I_j}{|\mathcal{S}_j|}}\right)F_{\mathbf{h}}^P. \quad (11)$$

Comparing equations (9) and (10) shows that the MMSE coefficients have an $I_{n_j/2}(\chi)/I_{n_j/2-1}(\chi)$ "weight" which reflects the uncertainty associated with using the Fourier coefficients of the known part to break down (or phase) the intensity datum. This produces the least biased of the estimates.

4. SIMULATIONS AND DISCUSSION

Two types of computational experiments were performed, the first to compare the performance of the MMSE and MAP estimates for realistic fiber diffraction data, and the second to assess the effect of the improved prior for simulated 2D point images.

To compare the performance of the MMSE estimate with the MAP estimates, synthetic fiber diffraction data I_j were calculated for the polymer mannan II [5]. The known part consists of the polymer backbone, while a set of water molecules (represented by oxygen atoms) are simulated as missing. The correlation coefficient

$$C = \frac{\int \hat{\varrho}^Q(\mathbf{r})\varrho^Q(\mathbf{r})d\mathbf{x}}{[(\int \varrho^Q(\mathbf{r})^2d\mathbf{x})(\int \hat{\varrho}^Q(\mathbf{r})^2d\mathbf{x})]^{1/2}}\,, \quad (12)$$

was used as a metric to quantify the quality of reconstructions. The data set comprises 71 Fourier coefficients $F_{\mathbf{h}}$. These were reduced to give 51 (data set I) and 41 (data set II) intensity data I_j, by a process of combining data with (approximately) the same cylindrical polar radius in Fourier space. The amount of missing structure was varied by amplifying the electron density of the water molecules, and quantified by the fraction of the total electron density ($\Delta\rho$) that is missing.

The resulting correlation coefficients are listed in Table 1. The performance of the MMSE estimate is better (larger C) than the MAP estimates throughout the range of missing structural information, and the improvement becomes more significant as the number of data decreases. Reconstructed images $\hat{\varrho}^Q(\mathbf{r})$ in a section containing two "missing" oxygen atoms, for data set I for $\Delta\rho = 0.05$, are shown as contour plots in Fig. 1. The MMSE reconstruction is seen to be the most faithful representation of the true image $\varrho^Q(\mathbf{r})$.

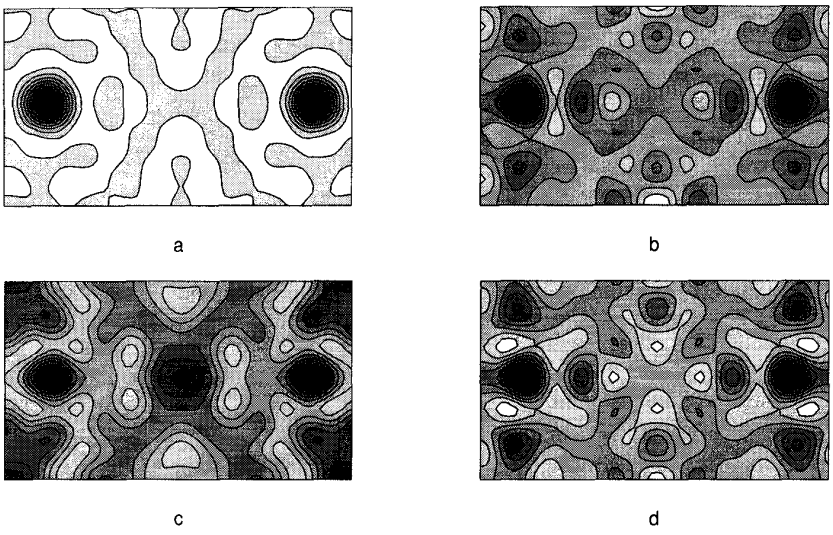

Figure 1: Contour plots of (a) the true missing electron density $\varrho^Q(\mathbf{r})$, and (b) MMSE, (c) MAP2, and (d) MAP1 estimates $\hat{\varrho}^Q(\mathbf{r})$ for mannan II. Darker regions indicate larger values.

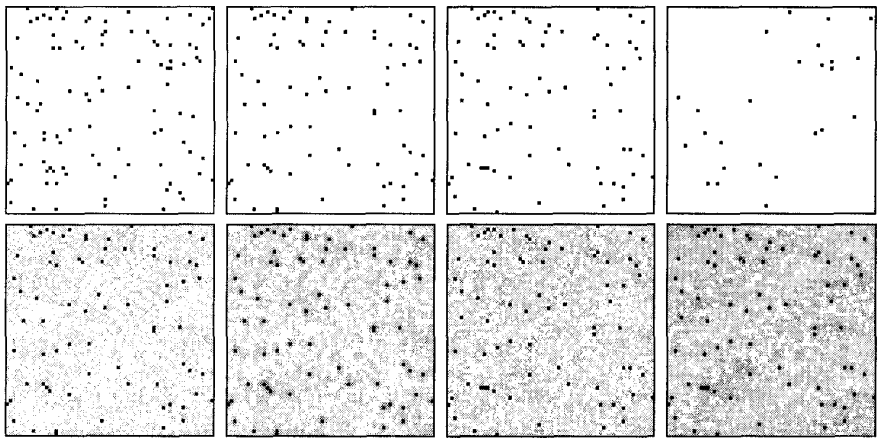

Figure 2: Top (L - R): The true image, $\varrho(\mathbf{r})$, the known part for $\sigma_P = 0.67$, the known part for $\sigma_P = 1.00$, and the missing part, for $|\mathcal{P}| = 75$. Bottom: The reconstruction of the full images with MMSE1 (left) and MMSE2 (next) for $\sigma_P = 0.67$, and the same for $\sigma_P = 1.00$ (right two images).

Table 1: Correlation coefficients for the three estimates for different amounts ($\Delta\rho$) of missing electron density and cylindrical overlap.

	Set I			Set II		
$\Delta\rho$	MMSE	MAP1	MAP2	MMSE	MAP1	MAP2
0.05	0.644	0.532	0.292	0.555	0.443	0.244
0.10	0.564	0.350	0.284	0.494	0.271	0.239
0.15	0.502	0.220	0.214	0.451	0.166	0.187

Table 2: Correlation coefficients for the estimated electron density for the two dimensional simulations with error in the partial structure.

$\|\mathcal{P}\|$	σ_P	MAP1	MAP2	MMSE1	MMSE2
60	0.67	0.301	0.207	0.313	0.341
60	1.00	0.179	0.147	0.185	0.213
75	0.67	0.379	0.336	0.387	0.428
75	1.00	0.261	0.236	0.268	0.306

To assess the effects of the improved prior, simulations were performed using two-dimensional random structures. A 64×64 grid was used, the structure consisted of 100 "point" atoms, of which 60 and 75 atoms were taken to be known. The known atoms were given random displacements with $\sigma_P = 0.67$ and 1.00 pixels. A cylindrical averaging over Fourier space produced a total of 353 intensity observations from the 2079 Fourier coefficients. The full image was reconstructed using both MAP estimates, and the MMSE estimate with the original (MMSE1) and improved (MMSE2) prior. The performance is compared using the correlation coefficients of the reconstruction with the true image and is tabulated in Table 2. The MMSE estimate with the original prior is only marginally better than the MAP1 estimate. However, incorporation of coordinate errors (D) in the improved prior significantly improves the correlation of the estimate.

The images obtained from the simulations for $|\mathcal{P}| = 75$ are shown in Fig. 2 to compare the effect of the improved prior on the MMSE estimate. The estimates incorporating the errors in the partial structure located more atoms at the correct positions (41 vs. 33 for $\sigma_P = 0.67$ and 32 vs. 23 for $\sigma_P = 1.00$) and exhibited a lower background noise. This leads to better interpretability of the reconstruction.

In practical applications of fiber diffraction analysis, the partial (polymer) structure is optimized against diffraction data from the full structure, and therefore contains coordinate errors. Incorporation of this feature into the prior as described here, and estimating the average coordinate errors from the data, can significantly reduce the noise and boost the accuracy of the estimated electron density. The Bayesian approach combined with full exploitation of prior knowledge gives optimal reconstructions, given the underdetermined nature of the problem.

5. REFERENCES

[1] R. P. Millane. *J. Opt. Soc. Am. A.*, 7:394–411, 1990.

[2] R. P. Millane. In N. W. Isaacs and M. R. Taylor, editors, *Crystallographic Computing 4: Techniques and New Technologies*, pages 169–186. Oxford: Oxford University Press, 1988.

[3] S. Baskaran and R. P. Millane. Proc. SPIE 3170:227–237, 1997.

[4] R. J. Read. *Acta Crystallogr.*, A42:140–149, 1986.

[5] R. P. Millane and T. L. Hendrixson. *Carbohydr. Polym.*, 25:245–251, 1994.