NONLINEAR RECONSTRUCTION PROBLEMS ARISING IN VIRAL STRUCTURE DETERMINATION FROM X-RAY AND ELECTRON MICROSCOPY DATA

Wen Gao, Yibin Zheng, and Peter C. Doerschuk

School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285 USA
{wgao, doerschu}@ecn.purdue.edu

ABSTRACT

We describe measurements, models, and algorithms for several signal reconstruction problems arising in the structural biophysics of so-called spherical viruses.

1. INTRODUCTION

In this paper we describe several signal reconstruction problems that arise during the determination of the 3D structure of so-called spherical viruses. Spherical viruses are viruses with a shell of protein (the capsid) surrounding an inner core of nucleic acid. The capsid is "crystalline" in the sense that it is constructed from many repetitions of the same polypeptides and the entire capsid is invariant under the rotational symmetries of the icosahedron. The icosahedron, as shown in Figure 1, is constructed from 20 equilateral triangles and has 60 rotational symmetries: a 5-fold axis where 5 triangles meet, a 3-fold axis through the center of each triangle, and a 2-fold axis at the midpoint of each edge between two triangles. A typical outer radius of the capsid is in the range 10^2–10^3Å.

2. MEASUREMENT PROCESSES

We consider three types of measurements: (1) x-ray diffraction from crystals of viral particles, (2) x-ray scattering from aqueous solutions of viral particles, and (3) cryo electron microscopy images of viral particles.

Let $\rho(\mathbf{x})$ [with Fourier transform $P(\mathbf{k})$] be the electron density in the crystal in real space or equivalently object space. $P(\mathbf{k})$ in reciprocal space or equivalently Fourier space is impulsive because $\rho(\mathbf{x})$ is periodic. The lattice of impulse locations is called the reciprocal lattice. The data in an x-ray crystal diffraction experiment, called intensities, are the magnitude-squared of the weights on the impulses of $P(\mathbf{k})$. One period of $\rho(\mathbf{x})$ is called the unit cell and occupies the volume S_u. Let $\rho_u(\mathbf{x})$ [with Fourier transform $P_u(\mathbf{k})$] be the electron density in the unit cell [i.e., $\rho_u(\mathbf{x}) = \rho(\mathbf{x})$ for $\mathbf{x} \in S_u$ and $= 0$ otherwise] which is therefore bandlimited in real space. The intensities are samples of $|P_u(\mathbf{k})|^2$.

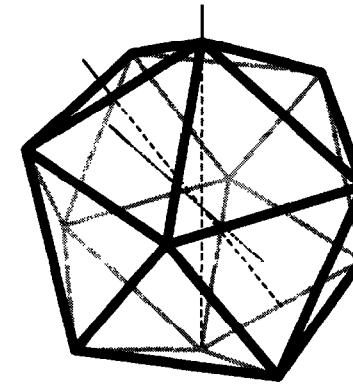


Figure 1: An Icosahedron. One symmetry axis of each type—2-, 3-, and 5-fold—is shown.

S_u is described in terms of a matrix $\mathbf{U} \in \mathcal{R}^{3\times3}$ by $S_u = \{\mathbf{U}\boldsymbol{\xi} : \boldsymbol{\xi} \in [0,1]^3\}$ where $\mathbf{U}$ is in general not an orthogonal matrix. Define $\kappa = (2\pi)^3/|\det(\mathbf{U})|$. By the definition of the unit cell, $\rho(\mathbf{x}) = \sum_{\mathbf{n}\in\mathcal{Z}^3} \rho_u(\mathbf{x} - \mathbf{Un})$ and therefore the Fourier transform is $P(\mathbf{k}) = P_u(\mathbf{k})\kappa\sum_{\mathbf{n}\in\mathcal{Z}^3}\delta(\mathbf{k} - 2\pi\mathbf{U}^{-T}\mathbf{n})$. Therefore the reciprocal lattice is the points $2\pi\mathbf{U}^{-T}\mathbf{n}$ for $\mathbf{n} \in \mathcal{Z}^3$ and the intensities (denoted by $F_\mathbf{n}$) are $F_\mathbf{n} = |P_u(2\pi\mathbf{U}^{-T}\mathbf{n})|^2\kappa^2$. Note that the matrix $\mathbf{U}$ is known before the processing described in this paper is performed.

The crystal is typically 50% by volume disordered solvent, with electron density ρ_0, and 50% by volume ordered viral particles. Therefore, the total electron density in the crystal is $\rho(\mathbf{x}) = \rho_0 + \tilde{\rho}(\mathbf{x})$ where $\tilde{\rho}(\mathbf{x})$ is the perturbation on the solvent electron density due to the viral particles. The term ρ_0 contributes only to the

Current address for Y.Z.: GE Corporate R&D, Room KWC 1303, P.O. Box 8, Schenectady, NY 12301. Supported by grants NSF DBI-9630497 and NSF DBI-9513594.

DC Fourier series coefficient and this coefficient is not measured because it coincides with the undiffracted x-ray beam. Therefore the data can be considered to be a function of $\tilde{\rho}(\mathbf{x})$. Because ρ_0 does not contribute to the data and the icosahedral symmetry of the viral particle determines the symmetries of $\tilde{\rho}(\mathbf{x})$, in the remainder of this paper we focus on $\tilde{\rho}(\mathbf{x})$ rather than $\rho(\mathbf{x})$ and for notational simplicity we drop the tilde from $\tilde{\rho}(\mathbf{x})$. Note, however, that the positivity condition $\rho(\mathbf{x}) \geq 0$ is equivalent to $\tilde{\rho}(\mathbf{x}) \geq -\rho_0$.

Let $\rho_v(\mathbf{x})$ [with Fourier transform $P_v(\mathbf{k})$] be the electron density perturbation on the solvent background created by the presence of one virus particle located at the origin and oriented in the standard orientation [1, 2]. The function $\rho_v(\mathbf{x})$ has icosahedral symmetry and therefore, by direct calculation, so does $P_v(\mathbf{k})$. A general formula for $\rho_u(\mathbf{x})$ for a unit cell containing Q virus particles is $\rho_u(\mathbf{x}) = \sum_{q=1}^{Q} \rho_v(\mathbf{T}_q^{-1}(\mathbf{x} - \mathbf{x}_q))$ where $\mathbf{x}_q$ ($\mathbf{x}_q \in \mathcal{R}^3$) is the position of the origin of the qth virus particle and $\mathbf{T}_q$ ($\mathbf{T}_q \in \mathcal{R}^{3\times 3}$, $\mathbf{T}_q^{-1} = \mathbf{T}_q^T$, $\det(\mathbf{T}_q) = +1$) is a rotation matrix that describes the orientation of the qth virus particle relative to the standard orientation. The corresponding Fourier transform is $P_u(\mathbf{k}) = \sum_{q=1}^{Q} \exp(-i\mathbf{k} \cdot \mathbf{x}_q) P_v(\mathbf{T}_q^{-1}\mathbf{k})$ and therefore the intensities are samples of

$$G(\mathbf{k}) = \kappa^2 \left| \sum_{q=1}^{Q} \exp(-i\mathbf{k} \cdot \mathbf{x}_q) P_v(\mathbf{T}_q^{-1}\mathbf{k}) \right|^2 .$$

Because the particles in the solution are randomly positioned and randomly rotated, the standard model [3] for the solution x-ray scattering is that the measured intensity is the spherical average of the magnitude-squared of the Fourier transform of the electron density in one particle. Therefore, in terms of the previous notation, the solution x-ray scattering [denoted by $I(k)$] is

$$I(k) = \frac{1}{4\pi} \int |P_v(\mathbf{k})|^2 \mathrm{d}\Omega'$$

where $\int \mathrm{d}\Omega'$ denotes integration over solid angles, $\mathrm{d}\Omega = \sin(\theta')\mathrm{d}\theta'\mathrm{d}\phi'$ in spherical coordinates, and $k = |\mathbf{k}|$. Notice that data is available continuously in k, that is, there is no sampling.

The cryo electron microscopy (EM) data is related to the 2D projection onto the object plane [denoted by $\sigma(x, y)$] of the 3D scattering density. Let $\sigma^i(x, y)$ be the image. Let $\Sigma(k, \phi')$ and $\Sigma^i(k, \phi')$ be the 2D Fourier transforms of $\sigma(x, y)$ and $\sigma^i(x, y)$ where polar coordinates are used for the transforms. Then, $\Sigma^i(k, \phi') = -\Sigma(k, \phi')A(k)\frac{2}{\lambda}f(k)\sin\chi(k)$ where $A(k)$ is the aperture function, λ is the electron wavelength, $f(k)$ is the atomic scattering factor for elastic scattering, and $\chi(k)$ is the phase shift due to spherical aberration and defocusing. The form of the phase shift is known: $\chi(k) = (2\pi/\lambda)(-C_s k^4/4 + \Delta f k^2/2)$ where C_s is the coefficient of spherical aberration and Δf is the deviation from Gaussian focus [4]. This theory can be elaborated to include the effects of specimen thickness (leading to varying levels of defocus), chromatic aberration, partial coherence, *etc.* Based on the projection slice theorem in 3D, $\Sigma(k_x, k_y) = P_v(\mathbf{R}_{\alpha,\beta,\gamma}\mathbf{k})|_{\mathbf{k}=(k_x,k_y,0)^T}$ where (α, β, γ) are Euler angles describing the projection and $\mathbf{R}_{\alpha,\beta,\gamma}$ is the corresponding rotation matrix.

Two central problems in cryo EM imaging are the unknown projection orientation and the sensitivity of the specimen to the electron beam. The goal is to reconstruct the 3D scattering density of the 3D specimen. The microscope produces an image that is closely related to the 2D projection $\sigma(x, y)$ as described above. Because the orientation of the 3D specimen on the stage of the microscope is not known, it follows that the image is related to an unknown-orientation 2D projection of the 3D specimen. This would not be a problem if the user could rotate the 3D specimen and take a series of images with known relative orientation, which is essentially what is done in medical imaging. However, this approach is not possible in cryo EM because of the sensitivity of the 3D specimen to the electron beam (and also due to technical problems with the range of achievable rotation). Therefore, taking multiple images of one 3D specimen in different orientations is replaced by taking one (or a very few) images of each of many identical 3D specimens where each specimen is in a random unknown orientation. Therefore the orientation parameters (α, β, γ) are not known.

3. VIRAL MODELS

The viral particle[1] has several characteristics:

1. Icosahedral constraint: $\rho(\mathbf{x})$ has icosahedral symmetry, that is, $\rho(\mathbf{R}_\beta^{-1}\mathbf{x}) = \rho(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{R}^3$ and $\beta \in \{0, \ldots, 59\}$.

2. Support constraint: $\rho(\mathbf{x}) = 0$ for $|\mathbf{x}| \leq R_-$ and $|\mathbf{x}| \geq R_+$.

3. Real-valued constraint: $\rho(\mathbf{x})$ is real.

4. Positivity constraint: $\rho(\mathbf{x}) \geq -\rho_0$ for all $\mathbf{x} \in \mathcal{R}^3$.

In addition, it is desirable to have a mathematical representation of $\rho(\mathbf{x})$ from which $P(\mathbf{k})$ can be computed analytically since this computation is a three-dimensional integral.

[1] In the remainder of this paper, $\rho(\mathbf{x})$ and $P(\mathbf{k})$ mean $\rho_v(\mathbf{x})$ and $P_v(\mathbf{k})$.

We have considered three different mathematical models for $\rho(\mathbf{x})$. All depend in an important way on icosahedral harmonics. Because of the icosahedral symmetry (a rotational symmetry) and the maximum radius for the region in which $\rho(\mathbf{x})$ may be non-zero, it is natural to use spherical coordinates in both real and reciprocal spaces. In order to easily compute the Fourier transform relating $\rho(\mathbf{x})$ to $P(\mathbf{k})$, it is natural to define real-space basis functions that are products of harmonic angular functions and spherical-Bessel radial functions. The angular function determines the rotational symmetry of the basis function. Spherical harmonics [5, Eq. 3.53], denoted by $Y_{l,m}(\theta,\phi)$ ($l \in \{0,1,\ldots\}$, $m \in \{-l,\ldots,+l\}$) are a complete orthonormal (CON) basis for L_2 functions on the sphere. However, $\rho(\mathbf{x})$ must have icosahedral symmetry and therefore we only need a CON basis for the subspace of icosahedrally symmetric L_2 functions on the sphere. Such a basis is provided by icosahedral harmonics [6, 7, 1, 2], denoted by $T_{l,n}(\theta,\phi)$ ($l \in \{0,1,\ldots\}$, $n \in \{0,1,\ldots,N_l-1\}$). Use of icosahedral rather than spherical harmonics means that any superposition of these basis functions has icosahedral symmetry and that the number of basis functions for each l is markedly decreased, specifically, N_l (for which formulas are known [6]) versus $2l+1$.

The first model, the so-called envelope model, is a piecewise constant model of $\rho(\mathbf{x})$:

$$\rho(r,\theta,\phi) = \begin{cases} \rho_c, & 0 \le r < \gamma^{\text{in}}(\theta,\phi) \\ \rho_s, & \gamma^{\text{in}}(\theta,\phi) \le r < \gamma^{\text{out}}(\theta,\phi) \\ 0, & \gamma^{\text{out}}(\theta,\phi) \le r \end{cases}.$$

Notice that this model does not allow overhanging regions on the surface of the virus or voids within the virus and is therefore restricted to low resolution. Mathematically, a piecewise constant model with a connected region is less restrictive while a convex region is more restrictive. The icosahedral symmetry of the virus implies that $\gamma^{\text{in}}(\theta,\phi)$ and $\gamma^{\text{out}}(\theta,\phi)$ must have icosahedral symmetry. Therefore, both can be expanded in an infinite sum of icosahedral harmonics and we use a finite truncation of the sum as the basis for computation:

$$\gamma^{\text{in}}(\theta,\phi) = \sum_{l=0}^{L^{\text{in}}} \sum_{n=0}^{N_l-1} \gamma^{\text{in}}_{l,n} T_{l,n}(\theta,\phi),$$

$$\gamma^{\text{out}}(\theta,\phi) = \sum_{l=0}^{L^{\text{out}}} \sum_{n=0}^{N_l-1} \gamma^{\text{out}}_{l,n} T_{l,n}(\theta,\phi).$$

In 2D, an example of this type of model is shown in Figure 2. The Fourier transform $P(\mathbf{k})$ and solution x-ray scattering $I(k)$ can be computed through the following

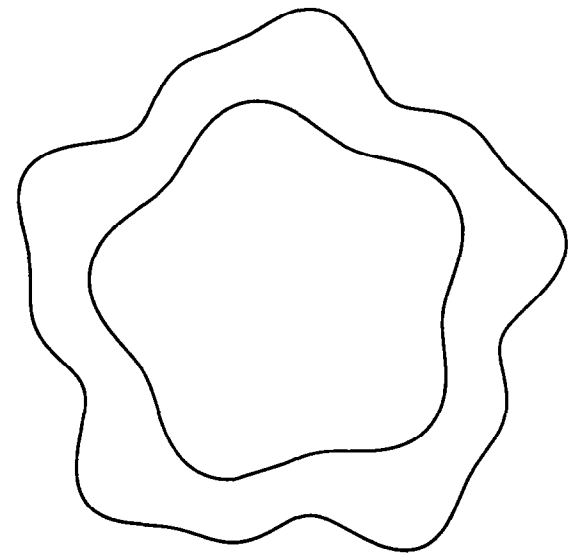


Figure 2: An example of an envelope model in 2D.

equations:

$$P(\mathbf{k}) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} (-i)^l a_{l,n}(k) T_{l,n}(\theta',\phi')$$

$$I(k) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} a^2_{l,n}(k)$$

$$a_{l,n}(k) = \frac{1}{k^3} \int \Big[\rho_s \mu_l(k\gamma^{\text{out}}(\theta,\phi)) + (\rho_c - \rho_s)\mu_l(k\gamma^{\text{in}}(\theta,\phi)) \Big] T_{l,n}(\theta,\phi) \mathrm{d}\Omega$$

$$\mu_l(x) = \int_0^x y^2 j_l(y) \mathrm{d}y.$$

The second model, the so-called orthonormal expansion model, is an orthonormal expansion of $\rho(\mathbf{x})$ in terms of icosahedral harmonics and spherical Bessel functions. There is no longer a piecewise-constant constraint. The definitions are:

$$\rho(r,\theta,\phi) = \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} \left[\sum_{p=1}^{\infty} c_{l,n,p} H_{l,p}(r) \right] T_{l,n}(\theta,\phi)$$

$$H_{l,p}(r) = \frac{y_l(\gamma_{l,p}R_+) j_l(\gamma_{l,p} r) - j_l(\gamma_{l,p}R_+) y_l(\gamma_{l,p} r)}{n_{l,p}}$$

$$n_{l,p} = \sqrt{\frac{R_-[j_l(\gamma_{l,p}R_-)]^2 - R_+[j_l(\gamma_{l,p}R_+)]^2}{2R_+\gamma^4_{l,p}R_-[j_l(\gamma_{l,p}R_-)]^2}}$$

where j_l and y_l are spherical Bessel functions of the first and second type, respectively, and where $\gamma_{l,p}$ for $p =$

$1, 2, \ldots$ are the roots of $j_l(\gamma R_-)y_l(\gamma R_+) - j_l(\gamma R_+)y_l(\gamma R_-) = 0$ and Sturm-Liouville theory guarantees that the minimum γ (i.e., $\gamma_{l,1}$) is finite.

The Fourier transform $P(\mathbf{k})$ and solution x-ray scattering $I(k)$ can be computed through the following equations:

$$P(\mathbf{k}) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} \sum_{p=1}^{\infty} c_{l,n,p} h_{l,p}(k)(-i)^l T_{l,n}(\theta', \phi')$$

$$I(k) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} \left[\sum_{p} c_{l,n,p} h_{l,p}(k) \right]^2$$

$$h_{l,p}(k) = \frac{R_+^2 H'_{l,p}(R_+) j_l(kR_+) - R_-^2 H'_{l,p}(R_-) j_l(kR_-)}{k^2 - \gamma_{l,p}^2}.$$

The third model, the so-called non-parametric model, is a generalization of the orthonormal expansion model. In this generalization, the function

$$A_{l,n}(r) = \sum_{p=1}^{\infty} c_{l,n,p} H_{l,p}(r)$$

is replaced by an unspecified function $A_{l,n}(r)$ satisfying $A_{l,n}(r) = 0$ for $r < R_-$ and $r > R_+$.

The Fourier transform $P(\mathbf{k})$ and solution x-ray scattering $I(k)$ can be computed through the following equations:

$$P(\mathbf{k}) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} (-i)^l a_{l,n}(k) T_{l,n}(\theta', \phi')$$

$$I(k) = 4\pi \sum_{l=0}^{\infty} \sum_{n=0}^{N_l-1} a_{l,n}^2(k)$$

$$a_{l,n}(k) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} r^2 A_{l,n}(r) j_l(kr) \mathrm{d}r$$

(spherical Hankel transform).

4. RECONSTRUCTION ALGORITHMS

In order to determine the 3D structure of the virus, it is necessary to estimate the following unknowns:

- Envelope model: ρ_s, ρ_c, $\{\gamma_{l,n}^{\text{in}}\}$, $\{\gamma_{l,n}^{\text{out}}\}$.
- Orthonormal expansion model: $\{c_{l,n,p}\}$ (assume R_- and R_+ are known).
- Non-parametric model: $\{A_{l,n}(r)\}$ (assume R_- and R_+ are known).

Two major techniques have been used: nonlinear weighted least squares for the envelope and orthonormal expansion models [8] and a set-projection algorithm for the nonparametric model [9]. The nonlinear least squares problems are solved by using the Levenberg-Marquart algorithm with analytical gradients.

5. NUMERICAL EXAMPLES

In Figure 3 we show numerical results for solution x-ray scattering data from Cowpea Mosaic Virus (CpMV) [10]. In the left and right hand columns of Figure 3 we show results based on the envelope and the orthonormal expansion models, respectively. The top image, which is common to both columns, shows the surface of the virus based on the known atomic-resolution structure for this virus. The lower four images show reconstructions based on synthetic and experimental data. Considering that the solution x-ray scattering data is 1D while a 3D reconstruction is desired, the reconstructions are surprisingly accurate, in large part due to the presence of the icosahedral symmetry.

6. REFERENCES

[1] Yibin Zheng and Peter C. Doerschuk. Explicit orthonormal fixed bases for spaces of functions that are totally symmetric under the rotational symmetries of a Platonic solid. *Acta Cryst.*, A52:221–235, 1996.

[2] Yibin Zheng and Peter C. Doerschuk. Symbolic symmetry verification for harmonic functions invariant under polyhedral symmetries. *Comput. in Phys.*, 9(4):433–437, July/August 1995.

[3] A. Jack and S. C. Harrison. On the interpretation of small-angle x-ray solution scattering from spherical viruses. *J. Mol. Bio.*, 99:15–25, 1975.

[4] O. Scherzer. The theoretical resolution limit of the electron microscope. *J. Appl. Phys.*, 20:20–29, January 1949.

[5] John David Jackson. *Classical Electrodynamics*. John Wiley, New York, 2 edition, 1975.

[6] Otto Laporte. Polyhedral harmonics. *Z. Naturforschg.*, 3a:447–456, 1948.

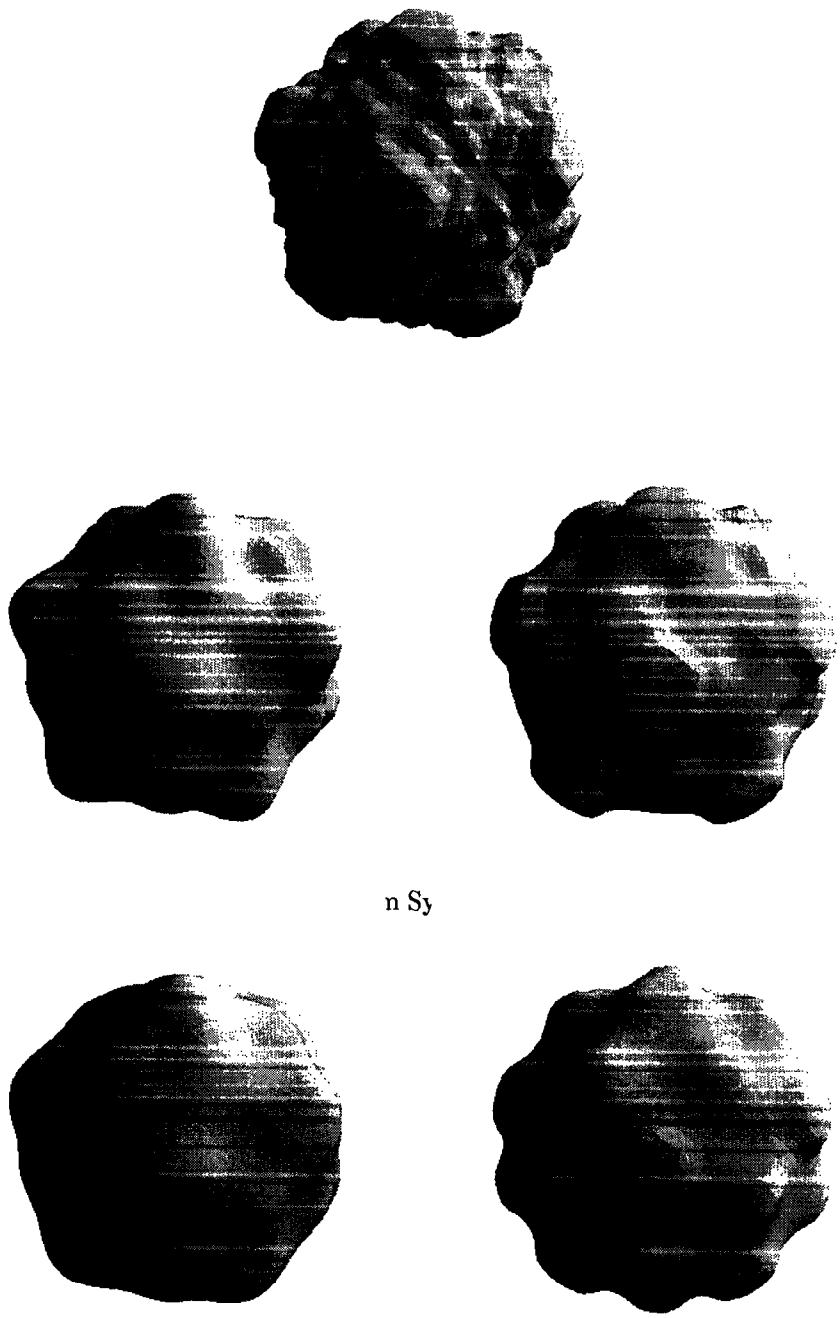
[7] Jacques Raynal. On a labeling for point group harmonics. II. Icosahedral group. *J. Math. Phys.*, 26(10):2441–2456, October 1985.

[8] Yibin Zheng, Peter C. Doerschuk, and John E. Johnson. Determination of three-dimensional low-resolution viral structure from solution x-ray scattering data. *Biophys. J.*, 69(2):619–639, August 1995.

[9] Yibin Zheng and Peter C. Doerschuk. Iterative reconstruction of three-dimensional objects from averaged Fourier-transform magnitude: solution and fiber x-ray scattering problems. *J. Opt. Soc. Am. A*, 13(7):1483–1494, July 1996.

[10] Zhongguo Chen, Cynthia V. Stauffacher, and John E. Johnson. Capsid structure and RNA packaging in comoviruses. *Seminars in Virology*, 1:453–466, 1990.

Envelope *Expansion Model*



n Sy

Exp

Figure 3: Numerical results for Cowpea Mosaic Virus. The parameters for the envelope model calculations are $L^{\text{in}} = L^{\text{out}} = 10$ which is 7 parameters and gives roughly 40Å resolution. The parameters for the orthonormal expansion model calculations are $L = 12$, $P_{0,0} = 3$, and $P_{6,0} = P_{10,0} = P_{12,0} = 2$ and the surface plotted is $\{\mathbf{x} : \hat{\rho}(\mathbf{x}) = 0.1 \max_{\mathbf{x}'} \hat{\rho}(\mathbf{x}')\}$.