Characterizing heterogeneity among virus particles by stochastic 3-D signal reconstruction

Nan Xu[a*], Yunye Gong[a*], Qiu Wang[a,b], Yili Zheng[c], and Peter C. Doerschuk[a,d]

[a]Electrical and Computer Engineering, Cornell Univ., Ithaca, NY, USA;
[b]Now with Siemens Corporate Research, Princeton, NJ, USA;
[c]Lawrence Berkeley National Laboratory, Berkeley, CA, USA;
[d]Meinig School of Biomedical Engineering, Cornell Univ., Ithaca, NY, USA

ABSTRACT

In single-particle cryo electron microscopy, many electron microscope images each of a single instance of a biological particle such as a virus or a ribosome are measured and the 3-D electron scattering intensity of the particle is reconstructed by computation. Because each instance of the particle is imaged separately, it should be possible to characterize the heterogeneity of the different instances of the particle as well as a nominal reconstruction of the particle. In this paper, such an algorithm is described and demonstrated on the bacteriophage Hong Kong 97. The algorithm is a statistical maximum likelihood estimator computed by an expectation maximization algorithm implemented in Matlab software.

Keywords: 3-D image reconstruction, statistical image processing, maximum likelihood estimation, expectation maximization algorithm, structural biology, virus, Hong Kong 97

1. INTRODUCTION

The geometric relationships among the components of a nanoscale biological machine such as a virus, ribosome, or enzyme are an important part of understanding the functioning of the machine and are the subject of structural biology. The most famous source for atomic-resolution information is x-ray crystallography. However, in recent years, single-particle cryo electron microscopy (cryo EM) has begun to provide information that is approaching atomic resolution due to improvements in both the image detectors and the computational tools.[1] In comparison with x-ray crystallography, cryo EM has the advantage of separately collecting information on each instance of the machine while x-ray crystallography collects information collectively on the large ensemble of instances that occur in the crystal. Therefore, cryo EM seems a likely basis for understanding the heterogeneity among the instances. In cases where the heterogeneity is discrete, i.e., that each instance belongs to a particular class of instances and all instances in a class can at least approximately be thought of as identical, then successful tools exist.[1] In this paper we describe a mathematical model and the resulting reconstruction algorithms that is focused on characterizing the continuous heterogeneity of particles within a discrete class.

In cryo EM, an aqueous specimen containing thousands of particles is flash frozen to cryogenic temperatures and imaged in the frozen state in an electron microscope. The reasons for freezing include reduction of the damage to the particles by the electron beam and the desire for a solid specimen. The image is basically a highly-noisy (SNR < 0.1) 2-D projection of the 3-D electron scattering intensity distribution of the particle. Primarily because of damage by the electron beam, only one projection is taken and the orientation of the projection direction is unknown and cannot be determined from the image because of low SNR. So, instead of reconstructing based on a full set of oriented projection images of a single particle, as is done in x-ray computed tomography in medical imaging, many images each of different instances of the particle and with different and unknown projection directions must be computationally combined to compute the reconstruction.[2] Standard

Further author information (Send correspondence to P.C.D.): N.X.: nx25@cornell.edu, Y.G.: yg326@cornell.edu, Q.W.: emmawq@gmail.com, Y.Z.: yzheng@lbl.gov, P.C.D.: pd83@cornell.edu.
N.X., Y.G., and P.C.D. are grateful to NSF 1217867 for funding.
* indicates authors who made an equal contribution to this work.

Image Reconstruction from Incomplete Data VIII, edited by Philip J. Bones, Michael A. Fiddy, Rick P. Millane,
Proc. of SPIE Vol. 9600, 96000F · © 2015 SPIE · CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2193791

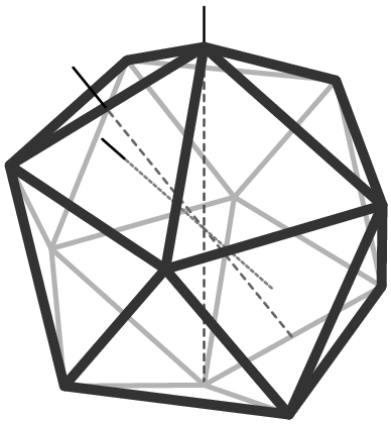


Figure 1. An icosahedron with one example of each type of rotational symmetry axis (2-, 3-, and 5-fold) shown.

approaches assume that the particles are either identical or are members of a small number of discrete classes such that all particles in a class are identical.

We have developed and demonstrated[3,4] an approach to reconstruction which includes both discrete classes and continuous variability of the particles within each class. Conditional on class membership, each particle is described as a linear combination of fixed basis functions where the weights in the linear combination are Gaussian random variables. The goal of the maximum likelihood (ML) reconstruction procedure is to determine the mean vector and covariance matrix of the vector of weights, which is a generalization of classical ML Gaussian mixture parameter estimation.[5]

Geometric symmetry of the particle is sometimes an important feature of the particle. Since the particle is of finite extent, such symmetries are point group symmetries [6, p. 26]. Therefore, there is at least one location in the particle that remains fixed under all of the symmetry operations and we take one such location as the origin of the coordinate system. Since the biological molecular constituents of the particle almost always have a fixed handedness, e.g., the biological amino acids are levo isomers, the point groups are of the first kind which contain only rotations. Because of our collaboration with Professor John E. Johnson (The Scripps Research Institute) and his colleagues, we have focused on viruses and for viruses the most important symmetry group is the icosahedral group with 60 operators (rotations of order 5, 3, and 2) and only one location which is fixed under all symmetry operations. The icosahedral group is the symmetry group of the icosahedron and a diagram of an icosahedron with examples of each type of symmetry axis is shown in Figure 1. An particle having this symmetry, however, need not look at all like an icosahedron! One of the contributions of this paper is to show how the same ideas and nearly the same software can also be used for particles with other symmetries, including no symmetry.

Our current Matlab software running on a desktop PC can solve problems with about 10^3 weights and achieve resolutions of 15–20Å on icosahedrally-symmetry (Figure 1) particles of diameter 432Å.[3] However, resolution twice as good, probably requiring $2^3 = 8$ times as many coefficients, or even four times as good, perhaps requiring $4^3 = 64$ times as many coefficients, is desirable. Furthermore, particles without symmetry require roughly 60 times as many coefficients since there are 60 symmetry operations in the icosahedral group. In order to achieve such gains in computational performance, another contribution of this paper is to present algorithmic improvements and software improvements.

2. THE MATHEMATICAL MODEL FOR THE PARTICLE AND IMAGING SYSTEM

We first describe the mathematical model for the particle[4] which is a critical component of the system because the model incorporates randomness in the description of the instances of the particle that belong to a particular discrete class. The class of the ith particle is denoted by $\eta_i \in \{1, \ldots, N_\eta\}$ and is a random variable with probability mass function q_i which is not known. Let $\mathbf{x} \in \mathbb{R}^3$ (x, θ, and ϕ in spherical coordinates) denote three

dimensional space ("real space") and $\mathbf{k} \in \mathbb{R}^3$ (k, θ', and ϕ' in spherical coordinates) the corresponding spatial frequency space ("reciprocal space"). The electron scattering intensity of the ith particle, denoted by $\rho_i(\mathbf{x})$, is described by a weighted linear combination of basis functions, specifically,

$$\rho_i(\mathbf{x}) = \sum_{\tau=1}^{N_c(\eta_i)} c_{i,\tau} \phi_\tau^{(\eta_i)}(\mathbf{x}) \tag{1}$$

where the unknown weights are denoted by $c_{i,\tau}$, the known basis functions are denoted by $\phi_\tau^{(\eta)}(\mathbf{x})$, and the number of weights and basis functions is $N_c(\eta)$. In x-ray crystallography or in other cryo EM algorithms, the weights $c_{i,\tau}$ are numbers and the goal of the computing is to determine numerical values for the $c_{i,\tau}$ from the data. Here we have generalized this idea: The $c_{i,\tau}$ are random variables and the goal of the computing is to determine the probability law for the random variables from the data. Let $c_i \in \mathbb{R}^{N_c(\eta_i)}$ be a random vector with components $c_{i,\tau}$. In order to make the calculations practical, we assume that the c_i ($i \in \{1 \ldots, N_v\}$) are independent random vectors and that, conditional on the value of the class η_i, c_i is distributed according to a Gaussian probability law with mean vector $\bar{c}^{(\eta_i)}$ and covariance matrix V_{η_i} where $\bar{c}^{(\eta)}$ and V_η are unknown. The goal of the computing is then to determine $\bar{c}^{(\eta)}$ and V_η ($\eta \in \{1, \ldots, N_\eta\}$) from the data. The mean vector $\bar{c}^{(\eta)}$ describes the nominal particle for Class η while the covariance matrix V_η describes the heterogeneity of the ensemble of instances of the particle in Class η. If there is no heterogeneity in a class then $V_\eta = 0$ and the model simplifies to the model we introduced and used previously.[7]

Let $\boldsymbol{\chi} \in \mathbb{R}^2$ denote two dimensional space in the images ("real space images") and $\boldsymbol{\kappa} \in \mathbb{R}^2$ the corresponding spatial frequency space ("reciprocal space images"). Let the real and reciprocal space images of the ith instance of the particle be denoted by $\sigma_i(\boldsymbol{\chi})$ and $\Sigma_i(\boldsymbol{\kappa})$, respectively. The theory[4] applies to a general linear imaging system where the linear transformation can depend on random variables so long as the random variables are independent of the random variables that describe the particle. We assume that the image is sampled. Therefore, if y_i^{NN} denotes a vector constructed from noise-free ("NN" for "no noise") samples of the ith image, there must be a matrix L such that

$$y_i^{\mathrm{NN}} = L_i(\theta_i, \eta_i) c_i \tag{2}$$

where θ_i are the additional random variables that enter into the linear transformation. Finally, we assume that the measured image, denoted by y_i, is corrupted by additive independent and identically distributed noise where the distribution of the noise is zero-mean Gaussian with covariance matrix Q. Let w_i denote the ith realization of the noise. Then the observed image, denoted by y_i, is described by

$$y_i = L_i(\theta_i, \eta_i) c_i + w_i. \tag{3}$$

The fact that c_i (conditional on the class label η_i) and w_i are both Gaussian is important for practical computation because a linear combination (as in Eq. 3) of Gaussian vectors is itself Gaussian with simple formulas for the mean vector and covariance matrix.

In the cryo EM application, the linear imaging system denoted by $L_i(\theta_i, \eta_i)$ is quite specific. In first-order image formation theory,[8–10] the reciprocal-space image $\Sigma_i(\boldsymbol{\kappa})$ is the product of three factors.

1. The 2-D Fourier transform of the ith projection image which, by the projection slice theorem, can be computed from the 3-D Fourier transform $P_i(\mathbf{k})$ of the particle $\rho_i(\mathbf{x})$ and the 3×3 rotation matrix $R_{\alpha_i,\beta_i,\gamma_i}$ that describes the projection direction which is parameterized by the Euler angles (α, β, γ).
2. The contrast transfer function $G(\kappa)$ which describes the electron-optical behavior of the microscope.
3. A complex exponential of the translation $\boldsymbol{\chi}_{0,i}$ of the projected location of the center of the ith particle from the center of the ith reciprocal-space image.

The resulting equation is

$$\Sigma_i(\boldsymbol{\kappa}) = \exp(-i2\pi \boldsymbol{\kappa}^T \boldsymbol{\chi}_{0,i}) G(\kappa) P^{(\eta_i)} \left(R^{-1}_{\alpha_i,\beta_i,\gamma_i} \begin{bmatrix} \boldsymbol{\kappa} \\ 0 \end{bmatrix} \right). \tag{4}$$

Discretize $\boldsymbol{\kappa}$ to the values $\boldsymbol{\kappa} \in \{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_{N_y}\}$. Let $\Phi_\tau^{(\eta)}(\mathbf{k})$ be the 3-D Fourier transform of $\phi_\tau^{(\eta)}(\mathbf{x})$ (Eq. 1). Then the variables in Eqs. 2 and 3 have the following forms:

1. The jth components of y_i^{NN} and y_i have the forms $\Sigma_i(\boldsymbol{\kappa}_j)$ and $\Sigma_i(\boldsymbol{\kappa}_j) + w_{i,j}$, respectively, where $w_{i,j}$ is the jth component of w_i.
2. The (j,τ)th element of the matrix $L_i(\theta_i, \eta_i)$ has the form

$$(L_i(\theta_i = (\alpha_i, \beta_i, \gamma_i, \boldsymbol{\chi}_{0,i}), \eta_i))_{j,\tau} = \exp(-i2\pi\boldsymbol{\kappa}_j^T \boldsymbol{\chi}_{0,i}) G(\kappa_j) \Phi_\tau^{(\eta_i)} \left(R_{\alpha_i,\beta_i,\gamma_i}^{-1} \begin{bmatrix} \boldsymbol{\kappa}_j \\ 0 \end{bmatrix} \right). \tag{5}$$

Conditional on the class $\eta = \eta'$, the statistics of $\rho(\mathbf{x})$ are Gaussian and so are completely described by the mean function $\bar{\rho}_{\eta'}(\mathbf{x})$ and covariance function $C_{\eta'}(\mathbf{x}, \mathbf{x}')$ which are defined by

$$\bar{\rho}_{\eta'}(\mathbf{x}) \doteq E[\rho(\mathbf{x}) | \eta = \eta'] = \sum_{\tau=1}^{N_c(\eta')} \bar{c}_\tau^{(\eta')} \phi_\tau^{(\eta')}(\mathbf{x}) \tag{6}$$

and

$$C_{\eta'}(\mathbf{x}, \mathbf{x}') \doteq E[[\rho(\mathbf{x}) - \bar{\rho}_{\eta'}(\mathbf{x})][\rho(\mathbf{x}') - \bar{\rho}_{\eta'}(\mathbf{x}')] | \eta = \eta'] = \sum_{\tau=1}^{N_c(\eta')} \sum_{\tau'=1}^{N_c(\eta')} (V_{\eta'})_{\tau,\tau'} \phi_\tau^{(\eta')}(\mathbf{x}) \phi_{\tau'}^{(\eta')}(\mathbf{x}') \tag{7}$$

where the second equalities of Eqs. 6 and 7 are both due to Eq. 1. Let $\hat{\bar{\rho}}_{\eta'}(\mathbf{x})$ and $\hat{C}_{\eta'}(\mathbf{x}, \mathbf{x}')$ be Eqs. 6 and 7 evaluated at the estimated values of $\bar{c}^{(\eta')}$ and $V_{\eta'}$ rather than the true values. For biological purposes, the natural quantities to visualize are $\hat{\bar{\rho}}_{\eta'}(\mathbf{x})$ and $\hat{C}_{\eta'}(\mathbf{x}, \mathbf{x}')$, especially the case $\hat{C}_{\eta'}(\mathbf{x}, \mathbf{x})$. The estimators used in this paper are maximum likelihood (ML) estimators (Section 4). For such estimators, if $y = f(x)$ and the ML estimate of x is $\hat{x}$ then the ML estimate of y is $f(\hat{x})$ [11, Theorem 7.2.10, p. 320]. Therefore, $\hat{\bar{\rho}}_{\eta'}(\mathbf{x})$ and $\hat{C}_{\eta'}(\mathbf{x}, \mathbf{x}')$ are ML estimates of $\bar{\rho}_{\eta'}(\mathbf{x})$ and $C_{\eta'}(\mathbf{x}, \mathbf{x}')$, respectively.

3. CHOICES FOR BASIS FUNCTIONS

The coordinate system used to describe the particle and the basis functions $\phi_\tau^{(\eta)}(\mathbf{x})$ can be chosen based on attractive mathematical properties and/or on attractive modeling properties.

Because of our collaboration with Professor Johnson, we have focused on virus particles with icosahedral symmetry. The symmetry implies that there is a unique location which is stationary under all of the symmetry operations and we select that location to be the origin of the coordinate system. Because the viruses are roughly spherical in shape, we use spherical coordinates.

Symmetry is a constraint on the electron scattering intensity $\rho_i(\mathbf{x})$ and therefore on the unknown weights $c_{i,\tau}$ but it is attractive to have an unconstrained maximization problem result from the maximum likelihood estimator. Therefore, selecting basis functions $\phi_\tau^{(\eta)}(\mathbf{x})$ such that $\rho_i(\mathbf{x})$ is symmetric for any choice of weights $c_{i,\tau}$ (Eq. 1) is attractive. Furthermore, the electron scattering intensity $\rho_i(\mathbf{x})$ is real valued so it is attractive to select basis functions $\phi_\tau^{(\eta)}(\mathbf{x})$ which are also real valued so that the weights $c_{i,\tau}$ can be real valued. Therefore, we have factored the basis functions $\phi_\tau^{(\eta)}(\mathbf{x})$ into a product of a function of x ("radial basis function") and a function of θ, ϕ ("angular basis function"). The symmetry constraint involves only the angular basis function.

For the angular basis functions we have derived so-called icosahedral harmonics[12,13] where each basis function is symmetric under all operators in the icosahedral group and the collection of all icosahedral harmonics is a complete orthonormal basis for icosahedrally-symmetric square-integrable functions on the surface of the sphere. The calculation is simplified by requiring each icosahedral harmonic to be a linear combination of spherical harmonics of a single degree. For the radial basis functions we solve a Sturm-Liouville problem on $[0, R_2]$ for spherical Bessel functions. The product of angular and radial basis functions is a complete orthonormal basis for square-integrable functions in the interior of the sphere of radius R_2. The index τ becomes a triple index (l, n, p) where $l \in \{0, 1, \ldots\}$ is the degree of the spherical harmonics that are linearly combined to give the

icosahedral harmonic, $n \in \{0, 1, \ldots, N_l - 1 \leq 2l\}$ is an index describing which icosahedral harmonic of degree l, and $p \in \{1, 2, \ldots\}$ is the index of the functions from the Sturm-Liouville problem where the Sturm-Liouville problem itself is indexed by l. Hence the basis functions, which are usually the same for all classes (i.e., all values of η), are $\phi_{\tau=(l,n,p)}(\mathbf{x}) = h_{l,p}(x) I_{l,n}(\theta, \phi)$.

Basis functions that are the product of these radial and angular basis functions have attractive properties. From the point of view of mathematics, they are real valued, have all the symmetries of the icosahedral group, and have simple symbolic formulas for their 3-D Fourier transforms. From the point of view of the application, for particles which fit more tightly into a sphere than into a rectangular parallelepiped, they do not represent the volume of real space where the electron scattering intensity of the particle is known to be zero and they also provide a smooth representation of $\rho_i(\mathbf{x})$ which can be evaluated at any value of $\mathbf{x}$ via Eq. 1.

A more sophisticated point of view is that the symmetry is a constraint on the statistics of $\rho_i(\mathbf{x})$ rather than on each individual realization of $\rho_i(\mathbf{x})$. Because we have assumed that the weights $c_{i,\tau}$ are Gaussian distributed, a complete description of the statistics of $\rho_i(\mathbf{x})$ is the mean function and the covariance function. We have presented the constraints on the mean and covariance functions [4, Eqs. 55 and 56] and have determined how these constraints control the statistics of the weights $c_{i,\tau}$ when using a set of angular basis functions in which each function transforms as one of the irreducible representations of the symmetry group.[14] However, we do not yet have complete software for this problem, specifically, in the estimation approach of Section 4, we do not have update algorithms for the covariance matrices V_η ($\eta \in \{1, \ldots, N_\eta\}$) which include the constraints on V_η that are implied by this more sophisticated view of symmetry.

While symmetry is important for some particles, many other important particles, such as ribosomes, lack symmetry. Without symmetries, the choice of spherical coordinates and harmonic angular basis functions is less compelling. Most software systems use rectangular coordinates and voxel basis functions which fill a rectangular parallelepiped and allow the use of many digital image processing ideas. However, especially for particles that are more closely fit by an sphere than a rectangular parallelepiped, it is still feasible to use spherical coordinates and harmonic angular basis functions and the natural angular basis functions are the so-called real-valued spherical harmonics [7, p. 1717, column 1] denoted by $\psi_{l,n}(\theta, \phi)$ and defined by

$$\psi_{l,n}(\theta, \phi) = \begin{cases} Y_{l,0}(\theta, \phi), & n = 0 \\ \sqrt{2}\Im Y_{l,(n+1)/2}(\theta, \phi), & n \in \{1, 3, 5, \ldots, 2l-1\} \\ \sqrt{2}\Re Y_{l,n/2}(\theta, \phi), & n \in \{2, 4, 6, \ldots, 2l\} \end{cases} \tag{8}$$

for $n \in \{0, \ldots, 2l\}$ where $Y_{l,m}(\theta, \phi)$ is the spherical harmonic of degree l and order m [15, Eq. 3.53, p. 99]. Changing the family of angular basis functions from icosahedral harmonics $I_{l,n}(\theta, \phi)$ to real-valued spherical harmonics $\psi_{l,n}(\theta, \phi)$ requires two changes to the software. First, the computation of $L_i(\theta_i, \eta_i)$ (Eq. 5) must be changed. Second, the integration rule used for integrating over the orientational nuisance parameters θ_i may need to be changed since the rule only has to integrate over one fundamental domain of the symmetry group. We have implemented the case of Eq. 8 in our software.

4. STATISTICAL ESTIMATORS

We compute a reconstruction by solving a maximum likelihood estimation problem for the unknown parameters which are the *a priori* class probabilities (q_η), the mean vector and covariance matrix for each class ($\bar{c}^{(\eta)}$ and V_η, respectively), and the covariance of the pixel noise vector (Q). We assume that Q is proportional to the identity matrix, i.e., $Q = \lambda^2 I_{N_y}$ where I_n is the $n \times n$ identity matrix. We have the theory for estimating the *a priori* probability distribution on the orientational Euler angles $(\alpha_i, \beta_i, \gamma_i)$ but successfully used a uniform distribution (i.e., Haar measure on the group SO_3) in our calculations. We have assumed that the origin offset $\boldsymbol{\chi}_{i,0}$ is uniformly distributed over a 2-D disk with known radius.

We have used an expectation maximization (EM) algorithm[5,16,17] to compute the maximum likelihood estimates where the nuisance parameters in the EM algorithm are the θ_i, i.e., the Euler angles $(\alpha_i, \beta_i, \gamma_i)$ and the origin offset $\boldsymbol{\chi}_{i,0}$, and the class label η_i.[3,4] We do not have practical procedures for updating all parameters (q_η, $\bar{c}^{(\eta)}$, V_η, and Q) simultaneously so we only update one or two at each iteration. Therefore, we are really using a generalized EM algorithm.

To save computation, when we update Q (really λ^2) we replace the EM update with a moment estimator as is described in the following. From Eq. 3 it follows that for every particle instance i, every component of the vector

$$\delta_i(\theta_i, \eta_i, c_i) = y_i - L_i(\theta_i, \eta_i)c_i \tag{9}$$

has variance λ^2 conditional on knowing the true values of θ_i, η_i, and c_i. Define $\theta = (\theta_1, \ldots, \theta_{N_v})^T$, $\eta = (\eta_1, \ldots, \eta_{N_v})^T$, and $c = (c_1, \ldots, c_{N_v})^T$. Define s^2 by

$$s^2(\theta, \eta, c) = \frac{1}{N_v N_y} \sum_{i=1}^{N_v} \sum_{j=1}^{N_y} (\delta_i(\theta_i, \eta_i, c_i))_j^2 \tag{10}$$

where $(\delta_i(\ldots))_j$ is the jth component of $\delta_i(\ldots)$. Let $\hat{\theta}$ be the current estimate of θ and likewise for $\hat{\eta}$. Let $\hat{\bar{c}}^{(\eta')}$ be the current estimate of $\bar{c}^{(\eta')}$ ($\eta' \in \{1, \ldots, N_\eta\}$). Then

$$\widehat{\lambda^2} = s^2(\hat{\theta}, \hat{\eta}, \hat{\bar{c}}^{(\hat{\eta})}) \tag{11}$$

is a natural moment estimator for λ^2 where $\hat{\bar{c}}^{(\hat{\eta})} = (\hat{\bar{c}}^{(\hat{\eta}_1)}, \ldots, \hat{\bar{c}}^{(\hat{\eta}_{N_v})})^T$. An alternative estimator is based on weighted averages over the uncertain values of (θ, η, c). Specifically, the estimator is

$$\widehat{\lambda^2} = \int_\theta \sum_\eta \int_c s^2(\theta, \eta, c) p(\theta, \eta, c) \mathrm{d}\theta \mathrm{d}c \tag{12}$$

where $p(\theta, \eta, c)$ is the current *a posteriori* probability distribution on these variables. We actually use a hybrid estimator defined by

$$\widehat{\lambda^2} = \int_\theta \sum_\eta s^2(\theta, \eta, \hat{\bar{c}}^{(\hat{\eta})}) p(\theta, \eta) \mathrm{d}\theta. \tag{13}$$

This estimator is straightforward to implement in our software. We have an excellent initial condition for λ^2 by estimating the value of λ^2 by the image sample variance in a region of the image where there are no particles. However, improving the estimate within the EM iterations appears to be important for experimental data problems.

5. NUMERICAL EXAMPLE

Some bacteriophage have a "head" which is a shell of protein (the "capsid") containing the phage's genome and a "tail" which is the component of the phage that recognizes a new host cell and through which the phage's genome is injected into the new host cell in order to initiate an infection. Typically, the head has icosahedral symmetry except in the region where the tail is attached where often a pentamer of the peptides that form the capsid are replaced by different peptides. Structural biology studies of the bacteriophage Hong Kong 97 (HK97) often use so-called "Virus Like Particles" because the wild-type bacteriophage has a flexible tail while VLPs that are only the head of the bacteriophage can be produced in bacteria that have been engineered to produce the capsid protein molecule. Because only the capsid proteins are produced, the VLP has icosahedral symmetry. While the head of the wild-type bacteriophage contains the viral genome, the VLP contains random cellular nucleic acid molecules. The algorithm of this paper was applied to a set of 1200 200×200 images of VLPs randomly selected from a larger collection provided by Professor J. E. Johnson and Dr. D. Veesler (both The Scripps Research Institute). Six example images are shown in Figure 2. The fact that the images in Figure 2(a) and Figure 2(b) are different in appearance emphasizes the need to include accurate and possibly even per-image contrast transfer functions in Eq. 4, i.e., $G_i(\kappa)$ or possibly $G_i(\boldsymbol{\kappa})$ which would allow for the inclusion of more detailed microscope aberrations. This is a challenging software engineering problem in our current Matlab software because it interferes with the matrix-matrix nature of the software which is important for the speed of the software.

A reconstruction using one class ($N_\eta = 1$) and 1060 basis functions based on icosahedral harmonics (all $l \leq 55$, n, and $p \leq 20$) was computed starting from an initial condition where no heterogeneity was allowed (i.e., $V = 0$).

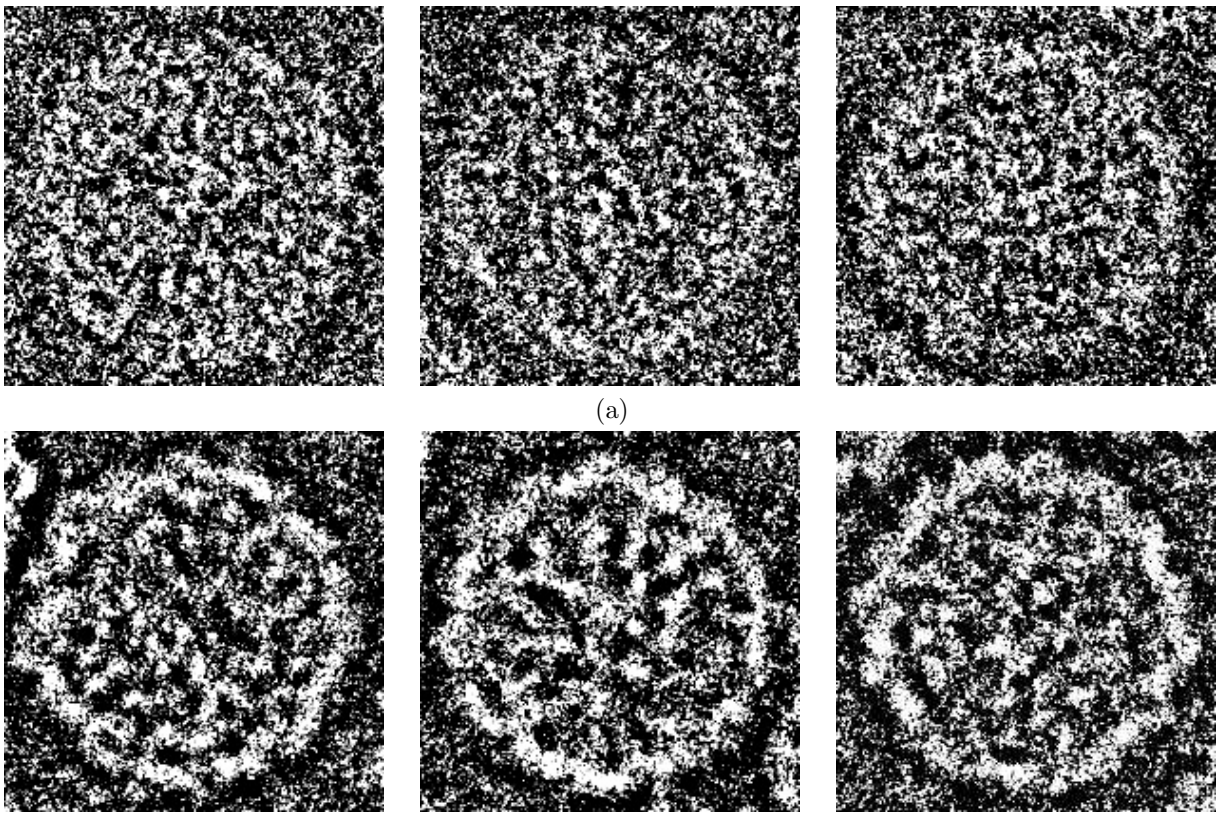


Figure 2. So-called boxed images extracted from the cryo EM micrographs of HK97. The images in Panel (a) versus in Panel (b) differ in the Contrast Transfer Function (CTF).

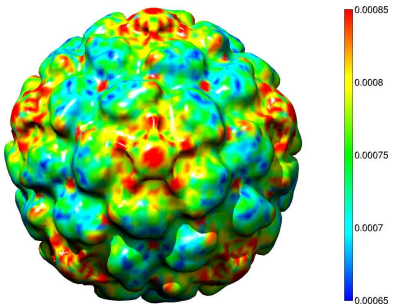


Figure 3. Surface from the mean $\hat{\bar{\rho}}(\mathbf{x})$ colored by the square root of the variance $\sqrt{\hat{C}(\mathbf{x},\mathbf{x})}$ for the HK97 reconstruction. Visualization by UCSF Chimera.[18]

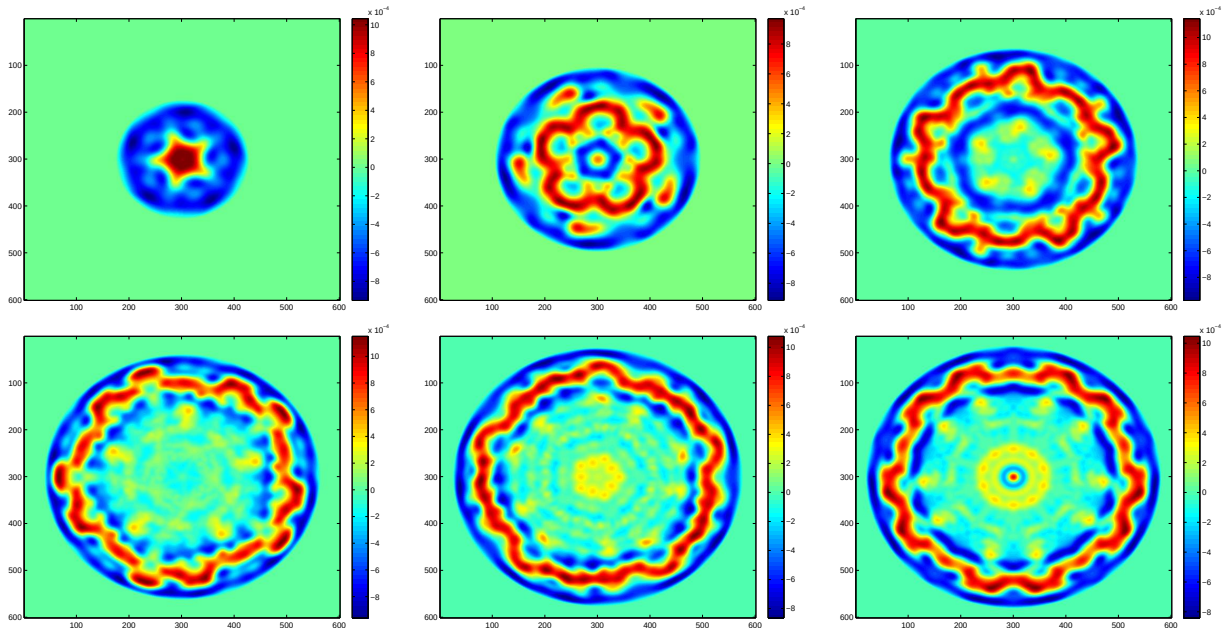
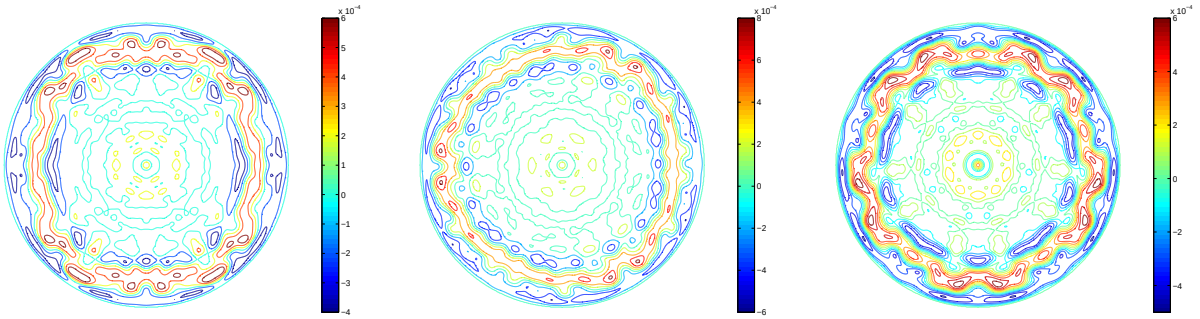


Figure 4. Serial cross sections of the mean $\hat{\tilde{\rho}}(\mathbf{x})$ in the reconstruction of HK97 where all sections are perpendicular to the z axis which is a 5-fold symmetry axis. The cube has a 1Å sampling interval and extends from -300Å to +300Å in each of the three rectangular coordinate directions. The slices are at $z \in \{-250, -200, -150, -100, -50, 0\}$Å and the other hemisphere can be filled in by icosahedral symmetry.



2-fold 3-fold 5-fold

Figure 5. Cross sections of the mean $\hat{\tilde{\rho}}(\mathbf{x})$ in the reconstruction if HK97. Each cross section is through the center of the particle and perpendicular to one of the three types of rotational symmetry axis.

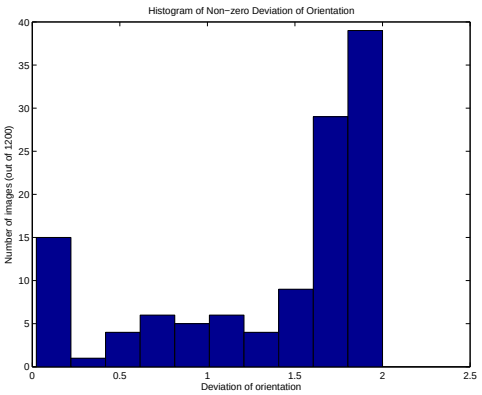


Figure 6. A typical histogram of $d(R_{\text{initial}}, R_{\text{final}})$ for the subset of images for which the orientation estimate changed during the heterogeneous HK97 reconstruction.

The mean vector $\bar{c}$ and covariance matrix V of the weights in the orthonormal expansion for $\rho(\mathbf{x})$ (Eq. 1) imply the mean function $\bar{\rho}(\mathbf{x})$ and the covariance function $C(\mathbf{x}_1, \mathbf{x}_2)$ of the electron scattering intensity (Eqs. 6 and 7). Figure 3 shows the external surface of the particle based on contouring the mean function $\bar{\rho}(\mathbf{x})$ and the surface is colored by the standard deviation function $\sqrt{C(\mathbf{x}, \mathbf{x})}$. The peak standard deviation is 2.23×10^{-3} while the average of the mean function is 4.88×10^{-4} so fluctuations can be substantial. The space-varying nature of the fluctuations is natural since in previous cases[3] it has been related to the functionality of the particle which is space-varying. While Figure 3 shows the surface of the particle, it is important to emphasize that the reconstructions, both mean and covariance, are 3-D. Therefore, in Figure 4 we show serial cross sections through half of the particle in order to demonstrate the 3-D nature of the reconstruction. Furthermore, in Figure 5, we show cross sections perpendicular to the three types of symmetry axis that occur in an icosahedrally-symmetric particle in order to demonstrate how use of these basis functions exactly and automatically enforces the icosahedral symmetry. In both Figure 4 and Figure 5 the capsid wall is clearly seen and the lack of structure in the internal cavity reflects the fact that the cavity is filled with random pieces of cellular nucleic acid.

In comparison with standard calculations, the calculations described in this paper extract additional information from the images because the continuous heterogeneity within each discrete class is characterized through the covariance matrices V_η. (In the HK97 example, $N_\eta = 1$ so $\eta = 1$ always). However, the calculations described in this paper also offer the possibility of improved quality for the nominal reconstruction because the heterogeneity of the particle is explicitly described in the calculations. One sense in which this can be observed is in the quality of the projection orientation estimates, i.e., the estimates of the Euler angles $(\alpha_i, \beta_i, \gamma_i)$ for the ith particle. Define the projection orientation of an image as the Euler angles of the integration abscissa for which the conditional probability density function is maximal. Following Huynh [19, Eq. 21], define a measure of difference between two rotations R_1 and R_2 by $d(R_1, R_2) = \|I_3 - R_1 R_2^T\|$ where $\|\cdot\|$ is the matrix norm induced by the Euclidean vector norm (the largest singular value). Since $R_2^T = R_2^{-1}$, d achieves its minimum possible value of 0 whenever $R_1 = R_2$. Since $\|I_3 - R_1 R_2^T\| \leq \|I_3\| + \|R_1 R_2^T\|$, both I_3 and $R_1 R_2^T$ are rotation matrices, and the singular values of a rotation matrix R are the eigenvalues of $RR^T = RR^{-1} = I$ which are all 1, it follows that $d(R_1, R_2) \leq 2$. When using the algorithm and software described in this paper, we first compute a reconstruction with no heterogeneity (i.e., $V = 0$) and then allow heterogeneity (i.e., $V \neq 0$). Let R_{initial} be the projection orientation estimate with $V = 0$ and R_{final} be the estimate with $V \neq 0$. The true projection orientations are not known for the HK97 data. Therefore, we cannot compare true and estimated projection orientations. Instead, we compare R_{initial} and R_{final}. The integration rule used in the HK97 example covers three fundamental domains of the icosahedral group for (α, β) and all of γ and contains 5000 abscissas so it is of only moderate resolution. At this resolution, only about 10% of the projection orientations change between R_{initial} and R_{final}. However, as is shown in the the histogram of $d(R_{\text{initial}}, R_{\text{final}})$ displayed in Figure 6, a majority of the projection orientations that do change are changed by a large amount and therefore contribute substantially differently to the nominal reconstruction.

Calculations on a range of genetic mutants of the HK97 particle are underway in the hope that comparison between the wild-type and the mutant reconstructions will lead to insights into the functioning of the particle.

6. SUMMARY AND DISCUSSION

In this paper we describe mathematical models, algorithms, and software which simultaneously compute a nominal reconstruction of the particle and characterize the heterogeneity of the instances of the particle from single-particle cryo electron microscopy images. The characterization of heterogeneity is complete in the sense that the mean function (the nominal reconstruction) and the complete covariance function of the electron scattering intensity are provided. While related work exists,[20–23] much of the related work is based on postprocessing with resampling which does not jointly estimate the nominal reconstruction of the particle and the heterogeneity of the instances of the particle, is computationally quite expensive, and does not usually provide the complete covariance function of the electron scattering intensity of the particle so that the characterization of heterogeneity is limited. The approach used in this paper is maximum likelihood estimation computed by a generalized expectation maximization algorithm and implemented in Matlab. Due to the characterization of heterogeneity and the maximum likelihood approach, these are large computations and we are presently investing substantial effort in improving our software system so that it will be more attractive to the structural biology community.

ACKNOWLEDGMENTS

We are grateful to Professor J. E. Johnson and Dr. D. Veesler (both The Scripps Research Institute) for the HK97 images and many discussions. N.X., Y.G., and P.C.D. are grateful to NSF 1217867 for funding.

REFERENCES

[1] Bai, X.-c., McMullan, G., and Scheres, S. H. W., "How cryo-EM is revolutionizing structural biology," *Trends in Biochemical Sciences* **40**, 49–57 (Jan. 2015).

[2] Jensen, G. J., ed., [*Cryo-EM, Parts A–C*], vol. 481–483 of *Methods in Enzymology*, Elsevier Inc. (2010).

[3] Wang, Q., Matsui, T., Domitrovic, T., Zheng, Y., Doerschuk, P. C., and Johnson, J. E., "Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps," *J. Struct. Biol.* **181**, 195–206 (Mar. 2013).

[4] Zheng, Y., Wang, Q., and Doerschuk, P. C., "3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object," *J. Opt. Soc. Am. A* **29**, 959–970 (June 2012).

[5] Redner, R. A. and Walker, H. F., "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review* **26**, 195–239 (Apr. 1984).

[6] Miller, Jr., W., [*Symmetry Groups and Their Applications*], Academic Press, San Deigo (1972).

[7] Doerschuk, P. C. and Johnson, J. E., "*Ab initio* reconstruction and experimental design for cryo electron microscopy," *IEEE Transactions on Information Theory* **46**, 1714–1729 (Aug. 2000).

[8] Erickson, H. P., "The Fourier transform of an electron micrograph—First order and second order theory of image formation," in [*Advances in Optical and Electron Microscopy (Volume 5)*], Barer, R. and Cosslett, V. E., eds., 163–199, Academic Press, London and New York (1973).

[9] Lepault, J. and Pitt, T., "Projected structure of unstained, frozen-hydrated T-layer of *bacillus brevis*," *EMBO J.* **3**(1), 101–105 (1984).

[10] Toyoshima, C. and Unwin, N., "Contrast transfer for frozen-hydrated specimens: Determination from pairs of defocused images," *Ultramicroscopy* **25**(4), 279–291 (1988).

[11] Casella, G. and Berger, R. L., [*Statistical Inference*], Duxbury, Wadsworth Group, Pacific Grove, CA, 2nd ed. (2002).

[12] Zheng, Y. and Doerschuk, P. C., "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).

[13] Zheng, Y. and Doerschuk, P. C., "Explicit computation of orthonormal symmetrized harmonics with application to the identity representation of the icosahedral group," *SIAM Journal on Mathematical Analysis* **32**(3), 538–554 (2000).

[14] Xu, N. and Doerschuk, P. C., "Reconstruction for stochastic 3-D signals with symmetric statistics in noise: electron microscopy of virus particles," in [*Proceedings of ICIP 2015, IEEE International Conference on Image Processing*], IEEE (27–30 September 2015).
[15] Jackson, J. D., [*Classical Electrodynamics*], John Wiley, New York, 2nd ed. (1975).
[16] McLachlan, G. J. and Krishnan, T., [*The EM Algorithm and Extensions*], Wiley-Interscience (1997).
[17] Bilmes, J. A., "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Tech. Rep. TR-97-021, Department of Electrical Engineering and Computer Science, University of California at Berkeley (Apr. 1998).
[18] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E., "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.* **25**(13), 1605–1612 (2004).
[19] Huynh, D. Q., "Metrics for 3D rotations: Comparison and analysis," *J. Math Imaging Vis.* **35**, 155–164 (2009).
[20] Penczek, Pawel A.and Kimmel, M. and Spahn, C. M. T., "Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images," *Structure* **19**, 1582–1590 (9 November 2011).
[21] Spahn, C. M. T. and Penczek, P. A., "Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM," *Current Opinion in Structural Biology* **19**, 623–631 (2009).
[22] Zhang, W., Kimmel, M., Spahn, C. M. T., and Penczek, P. A., "Heterogenneity of large macromolecular complexes revealed by 3D cryo-EM variance analysis," *Structure* **16**, 1770–1776 (2008).
[23] Penczek, P. A., Yang, C., Frank, J., and Spahn, C. M. T., "Estimation of variance in single-particle reconstruction using the bootstrap technique," *J. Struct. Biol.* **154**(2), 168–183 (2006).