

FROM HOMOGENEOUS TO HETEROGENEOUS:
STATISTICAL 3-D SIGNAL RECONSTRUCTION
OF MACROMOLECULAR COMPLEXES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Qiu Wang

May 2013

© 2013 Qiu Wang
ALL RIGHTS RESERVED

FROM HOMOGENEOUS TO HETEROGENEOUS: STATISTICAL 3-D
SIGNAL RECONSTRUCTION OF MACROMOLECULAR COMPLEXES

Qiu Wang, Ph.D.

Cornell University 2013

The structure and function of biological macromolecular complexes is currently a topic of great interest in biology. The primary contribution of this thesis is the mathematical description of a specific problem in this area of biology, the development of algorithms and software to solve the problem, and the demonstration of the relevance of the solution to biology.

The biology problem is to describe the three-dimensional structural heterogeneity of biological macromolecular complexes. The data is single-particle cryo electron microscopy images of individual instances of the complex and therefore the data contains information concerning the heterogeneity of the complex, although the information is usually ignored. Each image is a noisy 2-D projection of the 3-D electron scattering intensity of the particle modified by the electron optics of the microscope. This thesis focuses on developing statistical models, estimators for the parameters in the models, algorithms for determining the estimates, and computational implementations using high performance computing of the algorithms and demonstrates these results on biological problems where the complex is a virus.

The problem is treated as a stochastic signal in noise problem with the goal of estimating the statistics of the signal by a maximum likelihood estimator. The signal model includes both discrete and continuous heterogeneity, specifically, within each class of the discrete heterogeneity, the continuous heterogeneity

is described as Gaussian with unknown mean and covariance. The unknown *a priori* class probabilities and the unknown mean and covariance for each class are estimated by a maximum likelihood estimator which is solved by a generalized expectation-maximization algorithm which is implemented in parallel software. The software is demonstrated on experimental images from multiple types of viruses. Previously known biological results are reproduced and novel biological results are determined.

Different complexes have different spatial symmetry groups. Most of the work presented in this thesis concerns complexes that are roughly spherical in shape and especially the subset of such complexes which have icosahedral symmetry. The remainder of the work concerns complexes which have helical symmetry.

Evaluating the estimators requires substantial amounts of computation. Various algorithmic and software improvements to reduce computation are presented. For a fixed amount of computation, such improvements enable the achievement of higher spatial resolution in the estimated electron scattering intensity which will enable novel biological discoveries.

BIOGRAPHICAL SKETCH

Qiu Wang was born to a family of an electrical engineer and a computer engineer in Hangzhou, China. After graduated from high school, she moved to Hong Kong as a freshman of The Hong Kong University of Science and Technology with full fellowship. In 2007, She received her BS degree in the Honors Research Program of Electrical Engineering from The Hong Kong University of Science and Technology with first-class honors and the Academic Achievement Award. After a one-year exchange program during her senior year to Cornell University, she decided to pursue her Ph.D. study in the Department of Electrical and Computer Engineering at Cornell University. Advised by Prof. Peter C. Doerschuk, she worked on the research topic of statistical 3D signal reconstruction of macromolecular complexes, leading to the successful completion of this thesis.

I dedicate this thesis to my parents and my husband for their love and sacrifices.

ACKNOWLEDGEMENTS

I am extremely grateful to my thesis advisor, Prof. Peter C. Doerschuk, for introducing me to this exciting field of research. I was completely fascinated by the sophistication and beauty of applying tools like statistics, signal and image processing, mathematical programming, visualization and high performance computing onto scientific discoveries in macromolecular complexes and biophysics. I was constantly amazed by the breadth and depth of Peter's knowledge in electrical engineering, computer engineering, physics, mathematics and biomedical engineering. In addition to that, I was inspired and encouraged by his leadership in the face of what seemed insurmountable challenges. Peter has been instrumental to all of my professional accomplishments, from brainstorming the very beginnings of simple ideas that eventually matured into publications, to guiding me through the career decisions to help me carry on what we started at Cornell.

I thank my committee member, Prof. John (Jack) E. Johnson. Jack is a renowned structural biologist and virologist. I am grateful to him for agreeing to collaborate with me and introducing me to his virology problems, and teaching me everything I know about structural biology. In working with Jack during the course of our research, not only did I learn about structural biology and virology, but I also gained valuable insights into the method of questioning and the critical way of thinking in pure scientific discoveries. I thank my committee members, Prof. Lang Tong and Prof. Kevin Tang, for their thoughtful discussions. Dr. Tong is an expert in statistical inference, decisions and signal processing, future power energy systems and smart grids, information security, wireless communications and information theory. Dr. Tang is an expert in control and optimization of engineering networks such as the Internet and power

grids.

I thank all my collaborators in the Johnson Lab at The Scripps Research Institute, in particular, Dr. Tatiana Domitrovic, Dr. Tsutomu Matsui, Dr. David Veessler, Dr. Bradley Kearney and Dr. Chi-Yu Fu. I was always impressed by their devotedness as young scientists in structural biology. Their rich knowledge in structural biology and professional skill sets in experimental design have lead to several great applications and discoveries that we are all very much excited about.

I thank my collaborators in the Baker Lab at the University of California, San Diego, in particular, Prof. Timothy S. Baker and Dr. Jinghua Tang. Their and Prof. Johnson's expertise in understanding heterogeneous macromolecular complexes has been critical in the success of the research project.

I thank all the alumni of the Doerschuk Lab, in particular, Dr. Kang Wang, Dr. Yili Zheng, Dr. Seunghee Lee, Ms. Charlene Chen, Dr. Zhye Yin, Dr. Yibin Zheng and Dr. Cory Prust. I learned so much from all of them, either before they graduated from the lab, or afterwards through the works they have accomplished. I thank all my current lab mates, in particular, Ipek Ozil, Nan Xu and Nathan Cornelius. I learned so much from all of them and I hope the very best to everyone of them.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	vii
List of Figures	ix
1 Introduction	1
2 Statistical 3-D signal reconstruction of macromolecular complexes	6
2.1 Introduction	6
2.2 Statistical model	8
2.3 Relationship between V_η and $\rho(\mathbf{x})$	10
2.4 Estimation criterion	12
2.4.1 q as a function of \bar{c} , ${}_0V$, ${}_0Q$	13
2.4.2 \bar{c} , V , and Q as a function of \bar{c} , ${}_0V$, ${}_0Q$	14
2.4.3 Relationship with other results	19
2.5 Algorithm	20
2.6 Performance	20
2.7 Numerical results	23
2.8 Estimation of the <i>a priori</i> probability distribution on the nuisance parameters	25
2.9 Discussion	27
3 Applications on heterogeneous virus particles during their maturation over time	32
3.1 Heterogeneous <i>Nudaurelia Capensis</i> ω Virus (N ω V)	32
3.1.1 Introduction	32
3.1.2 Computational methodology	35
3.1.3 Results	53
3.1.4 Discussion	56
3.2 Other applications	71
3.2.1 Heterogeneity of NT procapsid, NT capsid and WT mature capsid of the N ω V particles	71
3.2.2 Heterogeneous <i>Hong Kong 97 Virus</i> (HK97)	72
4 Reciprocal space representations of helical-like objects with infinite period	73
4.1 Introduction	73
4.2 The case of spherically symmetric motifs	74
4.3 The case of general motifs	75
4.4 Discussion	76
4.5 Appendix	77

5	Computational performance optimization	84
5.1	Algorithm improvements	84
5.1.1	Implications of the matrix inversion lemma	84
5.1.2	Implications of Sylvester’s Determinant Theorem	86
5.1.3	Data-driven numerical integration rules	86
5.2	Software efficiency and supercomputing resources	89
6	Conclusion	94
6.1	Conclusions of this thesis	94
6.2	Future directions and challenges	95
	Bibliography	97

LIST OF FIGURES

2.1	Reconstruction of FHV from experimental images. Panel (a): Example boxed experimental images (same color map). Panel (b): Surface plot of the mean $\hat{\rho}_{\eta_0}(\mathbf{x})$ pseudo-colored by the variance $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x})$. The variance is highest near the 5-fold symmetry axes which is consistent with the idea that the binding of the particle to a new host cell and possibly later events concerning RNA translocation occur around a 5-fold axis [10, 13, 26]. Panel (c): Cross section of the mean through the center of the particle perpendicular to a 5-fold symmetry axis pseudo-colored by the variance. Panel (d): The RNA core after removal of the protein capsid. Surface plot of the mean pseudo-colored by the variance. The ordered dodecahedral RNA cage [20, 68] is detected and, as expected, the variance of the cage is low (blue) while that of the surrounding less well-ordered material is high (red). Visualizations in Panels (b–d) by UCSF Chimera [49].	29
2.2	Cross sections through the origin normal to 5-, 3-, and 2-fold symmetry axes of the estimated 3-D mean ($\hat{\rho}_{\eta=1}(\mathbf{x})$) function. The symmetries can clearly be seen in the cross sections, although the 5-fold symmetry is approximately a 10-fold symmetry and the 3-fold symmetry is approximately a 6-fold symmetry. The same color map is used in all images.	30
2.3	Fourier Shell Correlation (FSC) curve between $\hat{\rho}_{\eta=1}(\mathbf{x})$ computed with V constrained to $V = 0$ versus the optimal diagonal V . The independent variable k is the magnitude of the spatial frequency vector measured in Angstrom^{-1} . The pixel size of 4.7Angstrom implies a Nyquist frequency of $1/(2 \times 4.7) = 0.106\text{Angstrom}^{-1}$ which determined the upper limit of $k = 0.1\text{Angstrom}^{-1}$. Since the curve remains above $1/2$ for the entire range $k \in [0, 0.1]\text{Angstrom}^{-1}$, from the biological point of view the two mean structures are equivalent.	31
3.1	Algorithm flowcharts. The four parallel computations in Panel (a) are used to determine the performance of the algorithm, e.g., the error bars in Figure 3.6. The calculations contained in the red dotted-line box are expanded in Figure 3.2. . .	59
3.2	Algorithm flowcharts expanded. This is an expanded view of the red dotted-line box of Figure 3.1 which describes the maximum likelihood estimator.	60

3.3	The four time-resolved reconstructions. Panel A: Surface of each of the four reconstructions colored by the square root of the variance map (i.e., the standard deviation map) and displayed using the VIPERdb [77] convention. The same color map is used in all images. Panel B: The surface and a cross section perpendicular to a 2-fold axis of each of the four reconstructions colored by the standard deviation map. The surface and cross section visualizations at a particular time point share the same color map. Different color maps are used at different time points. Visualization by UCSF Chimera [49].	61
3.4	Part I of the resolution of the four time-resolved reconstructions as a function of k , which is the magnitude of the reciprocal-space frequency vector measured in Angstrom^{-1} . Fourier Shell Correlation (FSC) curves for comparing reconstructions from non-overlapping subsets containing 1200 images from the same data set. Based on these curves, the resolution of the four structures are approximately 21Angstrom. All FSC curves were computed using command <code>proc3d</code> in EMAN.	62
3.5	Part II of resolution of the four time-resolved reconstructions as a function of k , which is the magnitude of the reciprocal-space frequency vector measured in Angstrom^{-1} . Fourier Shell Correlation (FSC) curves between the 3 days reconstruction and each of the 3 minutes, 30 minutes, and 4 hours reconstructions for the nominal structures. Based on these difference curves, all the early structures agree with the capsid structure to approximately 24–33Angstrom. All FSC curves were computed using command <code>proc3d</code> in EMAN.	63
3.6	Part I of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. Panels A and B: Variance analysis around the cleavage sites of Subunits A, B, C and D that form the asymmetric unit of the NwV protein capsid. Both panels show the $T = 4$ surface lattice with the subunits' locations. The total volume occupied by each subunit is rendered as a mesh in Panel B. The variance was calculated over a smaller region, enclosing the cleavage site, which is shown as a solid volume within the subunit density. As is described in Section 3.1.2, the smaller region is essentially the region occupied by $C\alpha$ atoms within 10Angstrom of the active site. This is the same region analyzed by Matsui <i>et al.</i> [41] using difference maps.	64

3.7	Part II of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. The standard deviation for the regions displayed in Panel B of Figure 3.6 are plotted log-log as a function of time for each subunit. The plot demonstrates an overall reduction of variance as a function of time after maturation is initiated, with distinct kinetics between the variances of the B and C sites (high) and the A and D sites (low). Computational methods are described by Eqs. 3.33, 3.34, and 3.35. The capsid shell is defined to be the annulus with radius from 120Angstrom to 216Angstrom.	65
3.8	Part III of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. Panels D–E: Time variation of spherical averages. A cross section perpendicular to the 2-fold axis (Panel D) shows the location of the capsid shell relative to the center of the particle. The square root of the spherically-averaged variance map versus distance from the center of the particle was computed by Eqs. 3.37, 3.34, and 3.35 and is plotted in Panel E. The shaded region covers plus/minus one standard deviation. The inset plot shows a zoomed version of the plot including only the capsid shell region.	66
3.9	Ribbon diagrams of the four subunits at the four times colored by the square root of the variance map (i.e., the standard deviation map) with the asparagine at the self-catalytic site (Asn 570) shown as a ball-and-stick model. Each time point has its own color map analogous to the second row of Figure 3.3. For instance, red at the 3 minute time point is $14 \times 10^{-4} / 5.9 \times 10^{-4} = 2.4$ times higher than red at the 30 minute time point.	67
3.10	Surface of the 3 minute reconstruction colored by the square root of the variance map for a lower resolution reconstruction using 180 coefficients instead of 720 coefficients as was used in Figure 3.3 (the so-called Step 5 versus Step 7 of Ref. [83]). Visualization by UCSF Chimera [49].	68
3.11	Cross sections perpendicular to a 2-fold symmetry axis colored by the mean map (first row) or by the square root of the variance map (second row). The mean map is roughly binary while the variance map has substantial spatial variation. Visualization by UCSF Chimera [49].	69
3.12	Examples of images from the 3 minute image stack that were rejected by the algorithm of Section 3.1.2. These images were excluded from the reconstruction calculations.	70

5.1	Running time from a test case versus number of cores. The red bars are from parallel jobs using the local scheduler. The blue bars are from parallel jobs using the Matlab Distributed Computing Server (MDCS) and the TORQUE scheduler of the San Diego Supercomputer Center (SDSC).	92
5.2	Running time from a test case versus the number of cores in log-log scale, and compare with the reference line of the linear speed-up.	93

CHAPTER 1

INTRODUCTION

The structure and function of biological macromolecular complexes is currently a topic of great interest in biology. The primary contribution of this thesis is the mathematical description of a specific problem in this area of biology, the development of algorithms and software to solve the problem, and the demonstration of the relevance of the solution to biology.

Let x_i be the i th independent and identically distributed (i.i.d.) realization of a random vector from a Gaussian mixture probability density function (pdf). The pdf is

$$p(x | \{q_{\eta'}\}_{\eta'=1}^N, \{m_{\eta'}\}_{\eta'=1}^N, \{V_{\eta'}\}_{\eta'=1}^N) = \sum_{\eta=1}^N q_{\eta} \mathcal{N}(m_{\eta}, V_{\eta})(x) \quad (1.1)$$

where q_{η} is the *a priori* probability of the η th class ($\sum_{\eta=1}^N q_{\eta} = 1$, $q_{\eta} \geq 0$), m_{η} is the mean of the η th class, V_{η} is the covariance of the η th class, and \mathcal{N} is the multivariate Gaussian pdf. Let y_i be the i th measurement vector and suppose that

$$y_i = x_i. \quad (1.2)$$

In other words, the realizations of the Gaussian mixture random vector are directly observed. Suppose that the goal is to determine the values of q_{η} , m_{η} , and V_{η} from the measurements. Using the maximum likelihood approach, a standard algorithm is an expectation-maximization algorithm using the class labels of the measurements as the nuisance parameters. This iterative algorithm has an elegant set of update equations that involve nothing more complicated than numerical linear algebra. While the problem is usually posed in terms of q_{η} , m_{η} , and V_{η} the solution is really in terms of q_{η} , m_{η} , and V_{η}^{-1} which is natural because the Gaussian pdf is jointly concave in the mean and inverse covariance.

The biology problem is to describe the heterogeneity of biological macromolecular complexes. The data is electron microscopy images of individual instances of the complex. Fundamentally because of damage to the complex by the electron beam, only one image is recorded per complex. Therefore, the set of images contains information on the heterogeneity of the complex because it contains images of many different instances of the complex. Each image is essentially the 2-D projection of the 3-D electron scattering intensity of the complex (modified by the electron optics of the microscope) where the projection orientation is not known. Therefore, the image is a linear transformation of the electron scattering intensity of the complex, where the transformation has unknown parameters which are the projection orientation. In order to take advantage of the linear transformation and also in order to easily impose symmetry constraints, the electron scattering intensity is described as a weighted linear superposition of known basis functions. The goal of the calculation is to estimate the statistics of the weights from the data. Finally, the images are very noisy. The considerations of this paragraph motivates the model

$$y_i = L(z_i)x_i + w_i \tag{1.3}$$

where x_i are i.i.d. realizations of a random vector from a Gaussian mixture probability density with unknown parameters, L is the combination of the linear transformation from the weights to the 3-D electron scattering intensity distribution and the projection of the scattering intensity in the unknown orientation z_i to determine the image, and w_i is the noise (mean 0 and covariance V). The goal is to estimate the parameters of the Gaussian mixture pdf and the covariance of the noise w_i . These parameters then determine everything about the complex. A somewhat more general problem is also of interest. In particular, in some situations it may be desirable to use different basis functions to describe

different classes in the Gaussian mixture, e.g., different size or symmetry of the complex. In that case,

$$y_i = L(z_i, \eta_i)x_i + w_i \quad (1.4)$$

where the L matrix for the i th image depends not only on the unknown projection orientation for the i th image but also on the unknown class label η_i for the i th image. In Chapter 2 [87] a maximum likelihood solution for the problem of estimating q_i , m_i , Q_i , and V from the image data is described. In comparison with the situation of Eq. 1.2, there are several important challenges. First, there is now a continuous-valued nuisance parameter x_i as well as the discrete-valued nuisance parameter η_i . This leads to multi-dimensional integrals which cannot be evaluated symbolically in terms of standard functions and for which numerical evaluation for practical sized problems requires high-performance parallel computing. Second, the covariance of y_i conditional on the nuisance parameters is now a structured matrix, specifically,

$$\Sigma = L(z_i, \eta_i)V_{\eta_i}L^T(z_i, \eta_i) + Q. \quad (1.5)$$

Therefore, the change from optimizing with respect to V_{η} to optimizing with respect to V_{η}^{-1} is not a useful change.

Viruses are parasites. While viruses take advantage of molecular machinery of the infected host cell, the evolutionary pressures on the cell and its machinery are in the direction of hindering rather than aiding viruses. One evolutionary direction is simple viruses where the virus genome only encodes two or three peptides and where the production of virus progeny in an infected host cell has two steps: (1) an assembly step in which a complete virus assembles from constituent molecules in an essentially reversible non-covalent chemical reaction where the reversibility increases the yield of correctly formed virus particles fol-

lowed by (2) a maturation step in which the structure of the virus particle is reorganized in a non-reversible chemical reaction where non-reversibility leads to a virus particle that is durable outside of the cell. In Chapter 3 [82], the method of Chapter 2 is used to analyze time-resolved single-particle cryo electron microscopy (cryo-EM) images of Nudaurelia Capensis Omega Virus (N ω V) [41] which is a simple virus of the type described in the preceding sentences. Each virus particle contains 240 copies of the same capsid protein molecule in four different geometric positions. During maturation, each copy undergoes a autocatalytic cleavage reaction. Molecules in different geometries have different kinetics for this reaction which range from minutes to hours. Using time-resolved cryo-EM images (i.e., sets of images taken at specific times following the initiation of maturation), the method of Chapter 2 is able to demonstrate that capsid protein copies in different geometric positions have different time-varying trajectories of heterogeneity as measured by variance and that the heterogeneity tracks the kinetics previously measured by different methods [41].

Different viruses have different symmetries. The symmetry of N ω V is icosahedral symmetry. But other viruses and some artificial nanotechnology structures have helical symmetry. In Chapter 4, a method for describing the forward problem for electron microscopy of helical structures is derived which describes the helical symmetry in terms of two real parameters, a rotational offset and a translation offset between successive repeat units, rather than the more traditional one real parameter, the helical period, and two relatively-prime integer parameters which are u , the number of repeats per period, and v , where the rotational offset between repeat units is $2\pi v/u$. In solving inverse problems, it is potentially easier to optimize the two real-valued parameters rather than the one real and two integer-valued parameters.

The amount of computation required by the estimators described in this thesis is parameterized by several integers, e.g., number of images, number of pixels per image, number of basis functions, and so forth. For problems of biological interest, the amount of computation is moderately high which implies the need for efficient algorithms and high-performance software implementations. In Chapter 5, approaches and preliminary results are described for achieving faster computation. Example approaches include (1) the use of the Sherman-Morrison formula for matrix inversion, (2) the use of Sylvester's determinant theorem, and (3) the exploitation of known homogeneous reconstructions when computing heterogeneous reconstructions using the method of Chapter 2.

CHAPTER 2
STATISTICAL 3-D SIGNAL RECONSTRUCTION OF
MACROMOLECULAR COMPLEXES

2.1 Introduction

Single-particle cryo electron microscopy of multiple instances of a biological object in the 10^2 – 10^3 Angstrom spatial scale, such as a virus or a ribosome, is used to determine the object's 3-D structure, i.e., the spatial variation of the object's electron scattering intensity [21]. Each image is roughly a 2-D projection of the 3-D scattering intensity modified by the so-called contrast transfer function (CTF) of the microscope. Because the electron beam of the microscope rapidly damages the objects, most studies with spatial resolution goals of less than 25 Angstrom limit the dose in at least two ways. First, the dose per image is minimized, which implies computing reconstructions from images with low SNR, e.g., less than 0.1. Second, only one image is recorded per instance of the object, rather than a series of images at different projection orientations, so that reconstructions are computed by fusing information from one image of each of many instances of the object. Because the orientation in the microscope of the instances of the object are not controlled, the projection direction that results in a particular image is not known *a priori*. Because the image SNR is low, it is difficult from an individual image to determine either the projection direction that created that image or the projected location in the image of the center of the object. Most estimation and computation approaches assume that each instance of an object has identical 3-D structure. Exceptions include [16, 34, 57, 83] which assume that there are a few (e.g., 2–5) classes of objects and all instances

within a class are identical. However, there are important situations in which these assumptions are not valid. For example, at high spatial resolution (e.g., $< 5\text{Angstrom}$), most biological objects of this size have some element of flexibility and therefore instances in the same class are not identical even if each instance is composed of an identical number of identical chemical constituents. Furthermore, there exist many objects (e.g., Ref. [63]) where there is variability in the chemical constituents such as variability in the number of copies of a macromolecule. For the situation where each instance within a class has a different 3-D structure, this chapter (see also Refs. [81, 86]) describes a statistical model, estimator, and algorithm for determining the structure of each class as a statistical distribution which are based on generalizations of Gaussian mixture models, maximum likelihood (ML) parameter estimation, and generalized expectation maximization (EM) algorithms.

The problem considered in this chapter concerns data that is a linear transformation with structured stochasticity of a vector from a Gaussian mixture distribution with unknown parameters plus a second vector from a Gaussian distribution with known parameters. The goal is to estimate the unknown parameters of the Gaussian mixture distribution. By exploiting the Gaussian assumptions, the problem is equivalent to a Gaussian mixture problem with stochastically-structured mean and covariance matrices. Gaussian mixture problems with structured covariances have been studied in a wide range of application areas (e.g., remote sensing [5], speech [15, 60, 67], *etc.*) and from a large number of points of view (e.g., trees [67], constrained bases [60], a broad range of linear structures such as covariances arising from stationary time series [61], eigen decompositions with partial sharing between classes [5], subspaces with rank constraints where the subspaces are shared or unshared be-

tween classes [15], *etc.*). However, the problem considered in this chapter appears to be dominated by the complicated structure of the stochasticity of the linear transformation, which includes a projection from 3 to 2 dimensions, and therefore the approach taken is a development of the approach of Ref. [16].

The remainder of this chapter is organized in the following manner. The statistical model is described in Sections 2.2–2.3 and the estimation criterion and algorithm are described in Sections 2.4–2.5, including alternative forms of the algorithm suitable for different sizes of problem. The performance of the algorithm is discussed in Section 2.6. Numerical results based on Flock House Virus are presented in Section 2.7. Extensions for when the user desires to estimate the probability distribution for the nuisance parameters are described in Section 2.8. Finally, Section 2.9 contains a discussion.

2.2 Statistical model

Let $\eta \in \{1, \dots, N_\eta\}$ index the possible classes of an object where the number of classes, denoted by N_η , is known but the *a priori* probabilities of each class, denoted by q_η , are not known. Assume that the electron scattering intensity of an instance in the η th class is represented as a linear combination of $N_c(\eta)$ basis functions, denoted by $\phi_\tau^{(\eta)}(\mathbf{x})$, with support $S^{(\eta)} \subset \mathbf{R}^3$. Let $S = \cup_{\eta=1}^{N_\eta} S^{(\eta)}$. Therefore, the electron scattering intensity (denoted by $\rho(\mathbf{x})$) of an instance is

$$\rho(\mathbf{x}) = \sum_{\tau=1}^{N_c(\eta)} c_\tau \phi_\tau^{(\eta)}(\mathbf{x}) \quad (2.1)$$

where the unknown weights are denoted by c_τ . Since $\rho(\mathbf{x}) \in \mathbf{R}$, it is sufficient to consider systems where $c_\tau \in \mathbf{R}$ and $\phi_\tau^{(\eta)}(\mathbf{x}) \in \mathbf{R}$. Let $c \in \mathbf{R}^{N_c(\eta)}$ be a vector with components c_τ . Heterogeneity among instances within one class is described by

making c a Gaussian random vector with mean \bar{c}^η and covariance V_η . Reconstruction of the object is equivalent to estimation of \bar{c}^η and V_η for $\eta \in \{1, \dots, N_\eta\}$. Assume that the image is discretized and the samples are the components of a vector. Because the image formation process is linear, the unknown weight vector c and the image vector are related by a matrix denoted by L . The matrix depends on unknown parameters: the class of the instance (η) and the projection direction of the image and the location in the image coordinate system of the projection of the center of the object (the projection direction and center location are collectively denoted by θ).

The measurement noise in the image is described by an additive Gaussian zero-mean model ($N(\mu, \Sigma)(x)$ denotes the Gaussian pdf with mean μ and covariance Σ evaluated at location x). When the pixel noises are grouped into a vector, the covariance of this vector is denoted by Q where Q is to be estimated from the image data. The index $i \in \{1, \dots, N_v\}$ indicates which instance of the object. The possibility of recording several images at known relative projection directions, a so-called tilt series, is included by adding an index $j \in \{1, \dots, N_T\}$. Therefore the image formation model for the j th tilt of the i th instance is

$$y_{i,j} = L_{i,j}(\theta_i, \eta_i)c_i + w_{i,j} \quad (2.2)$$

where $\{\theta_i\}$, $\{\eta_i\}$, $\{c_i\}$, and $\{w_{i,j}\}$ are independent stochastic sequences; θ_i are i.i.d. with probability density function (pdf) $p(\theta)$ which is known (e.g., uniform over all rotations for projection orientation, i.e., Haar measure on the group $SO(3)$, times uniform on a disk of known radius for center location); η_i are i.i.d. with probability mass function (pmf) q_η which is unknown; c_i are independent but not identically distributed; for a fixed i , the pdf of c_i conditional on η_i is $N(\bar{c}^{\eta_i}, V_{\eta_i})$ where \bar{c}^{η_i} and V_{η_i} are unknown; and $w_{i,j}$ is i.i.d. (jointly in i and j) with pdf $N(0, Q_{i,j})$ where $Q_{i,j}$ is unknown. To include the possibility of a tilt series of

images, define $y_i = [y_{i,1}^T, \dots, y_{i,N_T}^T]^T$, $L_i(\theta_i, \eta_i) = [L_{i,1}^T(\theta_i, \eta_i), \dots, L_{i,N_T}^T(\theta_i, \eta_i)]^T$, $w_i = [w_{i,1}^T, \dots, w_{i,N_T}^T]^T$, and $Q_i = \text{diag}(Q_{i,1}, \dots, Q_{i,N_T})$. Then,

$$y_i = L_i(\theta_i, \eta_i)c_i + w_i \quad (2.3)$$

where w_i is i.i.d. with pdf $N(0, Q_i)$ and Q_i is unknown. Denote the conditional mean and covariance of y_i by

$$\mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}) \doteq E[y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i] \quad (2.4)$$

$$= L_i(\theta_i, \eta_i)\bar{c}^{\eta_i} \quad (2.5)$$

$$\Sigma_i(\theta_i, \eta_i, V_{\eta_i}, Q_i) \doteq \text{Cov}[y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i] \quad (2.6)$$

$$= L_i(\theta_i, \eta_i)V_{\eta_i}L_i^T(\theta_i, \eta_i) + Q_i \quad (2.7)$$

In this notation, the conditional pdf for y_i is

$$p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) = N(\mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}), \Sigma_i(\theta_i, \eta_i, V_{\eta_i}, Q_i))(y_i). \quad (2.8)$$

Collected in this paragraph is all of the notation used to describe the size of a problem. Some of this notation has not yet been used. The number of classes of object is N_η . The number of coefficients used to describe the electron scattering intensity of the η th class (Eq. 2.1) is $N_c(\eta)$. The number of objects imaged is N_v . The number of images taken of each object is N_T . The number of pixels in each image is N_y .

2.3 Relationship between V_η and $\rho(\mathbf{x})$

In this section we examine the relationship between V_η and $\rho(\mathbf{x})$ for the simplest possible structure for V_η . As described in Section 2.2, the support for the basis

functions $\phi_\tau^{(\eta)}(\mathbf{x})$ for the η th class is $S^{(\eta)}$. Let $I_S(\mathbf{x})$ be the indicator function for the set S . A simple model of heterogeneity in the η th class is that each instance's electron scattering intensity is the sum of a nominal intensity, denoted by $\rho_0^{(\eta)}(\mathbf{x})$, plus the restriction to $S^{(\eta)}$ of a zero-mean white perturbation, i.e.,

$$\rho(\mathbf{x}) = \rho_0^{(\eta)}(\mathbf{x}) + s_\eta(\mathbf{x})I_{S^{(\eta)}}(\mathbf{x}) \quad (2.9)$$

where $E[s_\eta(\mathbf{x})] = 0$ and $E[s_\eta(\mathbf{x})s_\eta(\mathbf{x}')] = \nu_\eta\delta(\mathbf{x} - \mathbf{x}')$. Assume that the basis functions are orthonormal, in which case,

$$c_\tau = \int_{S^{(\eta)}} [\rho_0^{(\eta)}(\mathbf{x}) + s_\eta(\mathbf{x})I_{S^{(\eta)}}(\mathbf{x})] \phi_\tau^{(\eta)}(\mathbf{x}) d^3\mathbf{x} \quad (2.10)$$

which implies

$$(\bar{c}^\eta)_\tau = E[c_\tau|\eta] \quad (2.11)$$

$$= \int_{S^{(\eta)}} \rho_0^{(\eta)}(\mathbf{x}) \phi_\tau^{(\eta)}(\mathbf{x}) d^3\mathbf{x} \quad (2.12)$$

$$(V_\eta)_{\tau,\tau'} = E[(c_\tau - E[c_\tau|\eta])(c_{\tau'} - E[c_{\tau'}|\eta])|\eta] \quad (2.13)$$

$$= \nu_\eta\delta_{\tau,\tau'}. \quad (2.14)$$

Therefore,

$$V_\eta = \nu_\eta I_{N_c(\eta)}. \quad (2.15)$$

The previous paragraph concerns the simplest structure for V . In the general case, by Eq. 2.1, it follows that the mean of the electron scattering intensity for a particular class is

$$\bar{\rho}_{\eta_0}(\mathbf{x}) \doteq E[\rho(\mathbf{x})|\eta = \eta_0] = \sum_{\tau=1}^{N_c(\eta_0)} \bar{c}_\tau^{\eta_0} \phi_\tau^{(\eta_0)}(\mathbf{x}) \quad (2.16)$$

and the autocorrelation is

$$r_{\eta_0}(\mathbf{x}, \mathbf{x}') \doteq E[[\rho(\mathbf{x}) - \bar{\rho}_{\eta_0}(\mathbf{x})][\rho(\mathbf{x}') - \bar{\rho}_{\eta_0}(\mathbf{x}')]| \eta = \eta_0] \quad (2.17)$$

$$= \sum_{\tau=1}^{N_c(\eta_0)} \sum_{\tau'=1}^{N_c(\eta_0)} (V_\eta)_{\tau,\tau'} \phi_\tau^{(\eta_0)}(\mathbf{x}) \phi_{\tau'}^{(\eta_0)}(\mathbf{x}'). \quad (2.18)$$

Let $\hat{\rho}_{\eta_0}(\mathbf{x})$ and $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x}')$ be Eqs. 2.16 and 2.18 evaluated at the estimated values of \bar{c} and V rather than the true values. For biological purposes, the natural quantities to visualize are $\hat{\rho}_{\eta_0}(\mathbf{x})$ and $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x}')$, especially the case $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x})$. The estimators used in this chapter are maximum likelihood (ML) estimators (Section 2.4). For such estimators, if $y = f(x)$ and the ML estimate of x is \hat{x} then the ML estimate of y is $f(\hat{x})$ [12, Theorem 7.2.10, p. 320]. Therefore, $\hat{\rho}_{\eta_0}(\mathbf{x})$ and $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x}')$ are ML estimates of $\bar{\rho}_{\eta_0}(\mathbf{x})$ and $r_{\eta_0}(\mathbf{x}, \mathbf{x}')$, respectively.

2.4 Estimation criterion

The goal of the estimation problem is to determine $\omega = \{q_\eta, \bar{c}^\eta, V_\eta, Q_{i,j} : \eta \in \{1, \dots, N_\eta\}, i \in \{1, \dots, N_v\}, j \in \{1, \dots, N_T\}\}$ for which no *a priori* pdfs are available. The parameters that describe the class and the projection orientation and coordinate system, $\Omega = \{\eta_i, \theta_i : i \in \{1, \dots, N_v\}\}$, are of less biological interest and have *a priori* pdfs. The measurement noise, $\{w_{i,j} : i \in \{1, \dots, N_v\}, j \in \{1, \dots, N_T\}\}$, is not of biological interest. The approach used in this chapter to estimate ω is maximum likelihood (ML) estimation. Once the estimate of ω , denoted by $\hat{\omega}$, is computed, it is sometimes useful to estimate Ω and that is done via $\hat{\Omega} = \arg \max_{\Omega} p(y|\hat{\omega}, \Omega)$ where $y = \{y_i : i \in \{1, \dots, N_v\}\}$. Define $q = \{q_\eta : \eta \in \{1, \dots, N_\eta\}\}$ and similarly for \bar{c} and V . Define $Q = \{Q_{i,j} : i \in \{1, \dots, N_v\}, j \in \{1, \dots, N_T\}\}$ and $Q_i = \{Q_{i,j} : j \in \{1, \dots, N_T\}\}$. By direct calculation, the log likelihood is

$$\ln p(y|\bar{c}, V, q, Q) = \sum_{i=1}^{N_v} \ln \left[\sum_{\eta_i=1}^{N_\eta} \int_{\theta_i} p(y_i|\theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) q_{\eta_i} p(\theta_i) d\theta_i \right] \quad (2.19)$$

where $p(y_i|\theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)$ is given in Eq. 2.8 and $p(\theta_i)$ is the *a priori* pdf on θ_i .

In order to compute the ML estimate, a generalized expectation-maximization (EM) algorithm is used. The sense in which the algorithm is a

generalized EM versus standard EM algorithm is that at any particular iteration, only a subset of the variables to be estimated are updated. Specifically, only (q, \bar{c}) , (q, V) , or (q, Q) are updated. As will be described in Section 2.4.1, the update for q is independent of the updates for the other variables but shares some of the same computations which motivates combining updates of q with the updating of each of the other variables. In the following paragraph, the expectation of the EM algorithm is derived and then, in Sections 2.4.1 and 2.4.2, the updates for q and for \bar{c} , V , and Q are derived.

The expectation in the EM algorithm is to compute

$$\begin{aligned}
& Q(\bar{c}, V, q, Q | \bar{c}, {}_0V, {}_0q, {}_0Q, y) \\
&= \int_{\theta} \sum_{\eta} [\ln p(y, \theta, \eta | \bar{c}, V, q, Q)] p(\theta, \eta | \bar{c}, {}_0V, {}_0q, {}_0Q, y) d\theta \quad (2.20) \\
&= \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_{\eta}} [\ln p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) + \ln p(\theta_i) + \ln q_{\eta_i}] p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, {}_0q, {}_0Q_i) d\theta_i \quad (2.21)
\end{aligned}$$

where \bar{c}, V, q, Q ($\bar{c}, {}_0V, {}_0q, {}_0Q$) are the new (old) values of the parameters. The conditional pdf on the nuisance parameters is

$$p(\theta_i, \eta_i | y_i, \bar{c}, V, q, Q) = \frac{p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) p(\theta_i) q_{\eta_i}}{\sum_{\eta'=1}^{N_{\eta}} \int_{\theta'} q_{\eta'} p(\theta') p(y_i | \eta', \theta', \bar{c}^{\eta'}, V_{\eta'}, Q_i) d\theta'}. \quad (2.22)$$

2.4.1 q as a function of $\bar{c}, {}_0V, {}_0Q$

Maximizing Eq. 2.21 with respect to q subject to the two constraints

$$\sum_{\eta'=1}^{N_{\eta}} q_{\eta'} = 1 \quad (2.23)$$

$$q_{\eta'} \geq 0 \quad \forall \eta' \in \{1, \dots, N_{\eta}\} \quad (2.24)$$

is equivalent to maximizing

$$Q_3(q|\bar{c}, {}_0V, \mathcal{A}, \mathcal{Q}, y) \doteq \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} [\ln q_{\eta_i}] p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{A}, \mathcal{Q}_i) d\theta_i \quad (2.25)$$

subject to the same two constraints. As is standard in the derivation of ML parameter estimates for Gaussian mixture models by EM, e.g., Refs. [8, 42, 53], the optimization problem is solved by ignoring Eq. 2.24, combining Eq. 2.23 with Eq. 2.25 by using a Lagrange multiplier, solving the resulting optimization problem, and verifying that the solution satisfies Eq. 2.24. The result is that

$$q_{\eta'} = \frac{1}{N_v} \sum_{i=1}^{N_v} \int_{\theta_i} p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{A}, \mathcal{Q}_i) d\theta_i \quad \forall \eta' \in \{1, \dots, N_\eta\} \quad (2.26)$$

where the computation of $p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{A}, \mathcal{Q}_i)$ is from Eq. 2.22.

2.4.2 \bar{c} , V , and Q as a function of \bar{c} , ${}_0V$, ${}_0Q$

Maximizing Eq. 2.21 with respect to \bar{c} , V and Q subject to

$$V_\eta = V_\eta^T \quad (2.27)$$

$$V_\eta > 0 \quad (2.28)$$

is equivalent to maximizing

$$\begin{aligned} & Q_1(\bar{c}, V, Q | \bar{c}, {}_0V, \mathcal{A}, \mathcal{Q}, y) \\ & \doteq \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} [\ln p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)] p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{A}, \mathcal{Q}_i) d\theta_i \quad (2.29) \\ & = -\frac{N_y}{2} \ln(2\pi) N_v - \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} \ln \det(\Sigma_i(\theta_i, \eta_i, V_{\eta_i}, Q_i)) p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{A}, \mathcal{Q}_i) d\theta_i \\ & \quad - \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}))^T \Sigma_i^{-1}(\theta_i, \eta_i, V_{\eta_i}, Q_i) (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i})) \times \\ & \quad \times p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{A}, \mathcal{Q}_i) d\theta_i \quad (2.30) \end{aligned}$$

(N_y is the number of pixels in an image) with respect to \bar{c} , V , and Q with the same two constraints.

\bar{c} as a function of V , Q , \bar{c} , 0V , 0Q

The gradient of Q_1 with respect to \bar{c} concerns only μ in the third term of Eq. 2.30.

The resulting gradient is

$$\begin{aligned} & \nabla_{\bar{c}^{\eta'}} Q_1(\bar{c}, V, Q | \bar{c}, {}^0V, \mathcal{Q}, {}^0Q, y) \\ &= - \sum_{i=1}^{N_y} \int_{\theta_i} L_i^T(\theta_i, \eta') \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) (L_i(\theta_i, \eta') \bar{c}^{\eta'} - y_i) p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}^0V_{\eta'}, \mathcal{Q}, {}^0Q_i) d\theta_i. \end{aligned} \quad (2.31)$$

Set

$$0 = \nabla_{\bar{c}^{\eta'}} Q_1(\bar{c}, V, q, Q | \bar{c}, {}^0V, \mathcal{Q}, {}^0Q, y) \quad (2.32)$$

for each $\eta' \in \{1, \dots, N_\eta\}$ to get a set of linear systems for the $\bar{c}^{\eta'}$ vectors for each $\eta' \in \{1, \dots, N_\eta\}$:

$$\begin{aligned} & \left[\sum_{i=1}^{N_y} \int_{\theta_i} L_i^T(\theta_i, \eta') \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) L_i(\theta_i, \eta') p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}^0V_{\eta'}, \mathcal{Q}) d\theta_i \right] \bar{c}^{\eta'} \\ &= \left[\sum_{i=1}^{N_y} \int_{\theta_i} L_i^T(\theta_i, \eta') \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) y_i p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}^0V_{\eta'}, \mathcal{Q}) d\theta_i \right] \end{aligned} \quad (2.33)$$

where $p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}^0V_{\eta'}, \mathcal{Q})$ is given by Eqs. 2.8 and 2.22. Please note that the $N_c(\eta') \times N_c(\eta')$ matrix on the LHS and the $N_c(\eta')$ vector on the RHS both depend on $V_{\eta'}$, which is not known.

V as a function of \bar{c} , Q , \bar{c} , 0V , 0Q

Define

$$N_i(y_i, \theta_i, \eta_i, \bar{c}^{\eta_i}) \doteq (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i})) (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}))^T \quad (2.34)$$

and use the trace operator to rewrite Eq. 2.30 in the form

$$\begin{aligned}
& \mathcal{Q}_1(\bar{c}, V, q, \mathcal{Q}|\bar{c}, {}_0V, \mathcal{Q}, {}_0\mathcal{Q}, y) \\
&= -\frac{N_y}{2} \ln(2\pi)N_v + \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} \ln \det(\Sigma_i^{-1}(\theta_i, \eta_i, V_{\eta_i}, \mathcal{Q}_i)) p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{Q}) d\theta_i \\
&\quad - \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_\eta} \text{tr} \left[\Sigma_i^{-1}(\theta_i, \eta_i, V_{\eta_i}, \mathcal{Q}_i) N_i(y_i, \theta_i, \eta_i, \bar{c}^{\eta_i}) \right] p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{Q}) d\theta_i.
\end{aligned} \tag{2.35}$$

It is not possible to follow the standard derivation of ML parameter estimates for Gaussian mixture models by EM, which involves computing derivatives with respect to Σ^{-1} , because it is necessary to account for the dependence of Σ on V . If f is a scalar-valued function of the matrix X , let $\partial f / \partial X$ be the matrix of partial derivatives of f with respect to the elements of X . Eq. 2.35 implies

$$\begin{aligned}
& \frac{\partial \mathcal{Q}_1(\bar{c}, V, q, \mathcal{Q}|\bar{c}, {}_0V, \mathcal{Q}, {}_0\mathcal{Q}, y)}{\partial V_{\eta'}} \\
&= \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \frac{\partial \ln \det(\Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, \mathcal{Q}_i))}{\partial V_{\eta'}} p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \\
&\quad - \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \frac{\partial \text{tr} \left[\Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, \mathcal{Q}_i) N_i(y_i, \theta_i, \eta', \bar{c}^{\eta'}) \right]}{\partial V_{\eta'}} p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i.
\end{aligned} \tag{2.36}$$

For full-rank square matrices X , let R be the function that is matrix inversion, i.e., $R(X) = X^{-1}$. Using the chain rule and a simplified notation and without imposing any constraints on the structure of the matrices Q or $V_{\eta'}$

$$\frac{\partial \ln \det(R(\Sigma(V)))}{\partial V} = - \left[L^T (\Sigma^{-1} + \Sigma^{-T}) L - \text{diag} (L^T \Sigma^{-1} L) \right] \tag{2.37}$$

$$\frac{\partial \text{tr}(R(\Sigma(V))N)}{\partial V} = - \left[L^T (\Sigma^{-1} N \Sigma^{-1} + \Sigma^{-T} N \Sigma^{-T}) L - \text{diag} (L^T \Sigma^{-1} N \Sigma^{-1} L) \right] \tag{2.38}$$

Specializing Eqs. 2.37 and 2.38 to the case where Q and $V_{\eta'}$ are transpose symmetric and applying the results to Eq. 2.36 gives

$$\frac{\partial \mathcal{Q}_1(\bar{c}, V, q, \mathcal{Q}|\bar{c}, {}_0V, \mathcal{Q}, {}_0\mathcal{Q}, y)}{\partial V_{\eta'}} = 2S(y, \eta', \bar{c}^{\eta'}, V_{\eta'}, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q})$$

$$- \text{diag} \left(S(y, \eta', \bar{c}^{\eta'}, V_{\eta'}, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \right) \quad (2.39)$$

where

$$\begin{aligned} & S(y, \eta', \bar{c}^{\eta'}, V_{\eta'}, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \\ & \doteq \frac{1}{2} \sum_{i=1}^{N_y} \int_{\theta_i} M_i(y_i, \theta_i, \eta', \bar{c}^{\eta'}, V_{\eta'}) p(\theta_i, \eta' | y_i, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \end{aligned} \quad (2.40)$$

$$\begin{aligned} & M_i(y_i, \theta_i, \eta', \bar{c}^{\eta'}, V_{\eta'}) \\ & \doteq L_i^T(\theta_i, \eta') \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, \mathcal{Q}_i) N_i(y_i, \theta_i, \eta', \bar{c}^{\eta'}) \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, \mathcal{Q}_i) L_i(\theta_i, \eta') \\ & \quad - L_i^T(\theta_i, \eta') \Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, \mathcal{Q}_i) L_i(\theta_i, \eta') \end{aligned} \quad (2.41)$$

where $p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, \mathcal{Q}_i)$ is given in Eq. 2.8. Setting the derivative given by Eq. 2.39 equal to zero implies

$$0 = S(y, \eta', \bar{c}^{\eta'}, V_{\eta'}, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \quad (2.42)$$

which, however, cannot be solved symbolically for the value of $V_{\eta'}$. Note that Eq. 2.42 depends on $\bar{c}^{\eta'}$, which is not known.

Rather than solve the nonlinear Eq. 2.42 for the new value of $V_{\eta'}$, we maximize \mathcal{Q}_1 directly. We consider two structures for V : (1) $V_{\eta} = v_{\eta} I_{N_c(\eta)}$ (Eq. 2.15) and (2) $V_{\eta} = \text{diag}(v_{\eta})$.

The case of $V_{\eta} = v_{\eta} I_{N_c(\eta)}$: Updating V_{η} , really updating v_{η} , is done by the Matlab [40] function `fminbnd` with a limit of 8 iterations and typically 8 iterations are used. Each iteration has nearly the computation complexity of one update of \bar{c} .

The case of $V_{\eta} = \text{diag}(v_{\eta})$: Updating V_{η} , really updating v_{η} , is done by a Newton's method with symbolic formula for the first and second derivatives because no line search is required. The first and second derivatives of the scalar \mathcal{Q}_1 with

respect to the components of the vector v_η can be written using standard matrix-vector notation. The gradient (a $N_c(\eta)$ -component vector) is a special case of Eq. 2.39, specifically

$$\begin{aligned} & \nabla_{v_\eta} \mathcal{Q}_1(\bar{c}, V, q, \mathcal{Q}|\bar{c}, {}_0V, \mathcal{Q}, y) \\ &= \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \text{diag}(M_i(y_i, \theta_i, \eta, \bar{c}^\eta, V_\eta)) p(\theta_i, \eta | y_i, \bar{c}, {}_0V, \mathcal{Q}, y) d\theta_i \end{aligned} \quad (2.43)$$

where diag indicates extracting the diagonal of a matrix to construct a vector. The Hessian (a symmetric $N_c(\eta) \times N_c(\eta)$ -element matrix) is

$$\frac{\partial^2 \mathcal{Q}_1(\bar{c}, V, q, \mathcal{Q}|\bar{c}, {}_0V, \mathcal{Q}, y)}{\partial v_\eta^2} = \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} (-2D_i \circ E + E \circ E) d\theta_i \quad (2.44)$$

where \circ is the Hadamard product (entry-by-entry product) and

$$D = L_i^T(\theta_i, \eta) \Sigma_i^{-1}(\theta_i, \eta, V_\eta) N_i(y_i, \theta_i, \eta, \bar{c}^\eta) \Sigma_i^{-1}(\theta_i, \eta, V_\eta) L_i(\theta_i, \eta) \quad (2.45)$$

$$E = L_i^T(\theta_i, \eta) \Sigma_i^{-1}(\theta_i, \eta, V_\eta) L_i(\theta_i, \eta). \quad (2.46)$$

Up to 8 Newton steps are taken, typically all 8 are used, and each step has nearly the computation complexity of one update of \bar{c} .

An alternative to the method of the previous paragraph is to approximate \mathcal{Q}_1 so that the approximation is concave in V and then use convex optimization methods. The log of the Gaussian pdf is jointly concave in the mean and inverse covariance and linear combinations of concave functions are concave so the expectation integral makes \mathcal{Q}_1 a concave function of \bar{c}^η and Σ_i^{-1} . However, we seek to optimize \mathcal{Q}_1 with respect to \bar{c}^η and V_η and Σ_i^{-1} is a complicated function of V_η . Since if $f(x)$ is concave in x then $g(y) = f(a + By)$ is concave in y , a linear approximation for Σ_i^{-1} as a function of V_η would lead to a convex optimization problem. Iteration application of a linearization of Σ_i^{-1} as a function of V_η around the current value of V_η followed by solution of the resulting convex

optimization problem provides a complete algorithm for updating V_η . For the case of $V_\eta = \text{diag}(v_\eta)$, the linearization of Σ_i^{-1} as a function of V_η around V_{η_0} has, in abbreviated notation, the simple form $\Sigma^{-1} = (LV_0L^T + Q)^{-1} - T(V - V_0)T^T$ where $T = Q^{-1}L(L^TQ^{-1}L + V_0^{-1})^{-1}V_0^{-1}$. This approach has not been implemented but will be reconsidered for larger problems.

Q as a function of $V, \bar{c}, \bar{\sigma}, {}_0V, {}_0Q$

Only the case of $Q_{i,j} = \lambda I_{N_y}$ is considered so there is only a single scalar parameter. An initial condition is available by computing the sample variance of the image in an annulus outside of the virus particle [83, Section 2.8]. Simple formulas are available for the first and second derivatives of Q_1 in the case of $V = 0$ (not shown) but the corresponding formulas for $V \neq 0$ are complicated. Therefore, numerical optimization was done using Matlab [40] function `fminbnd` with a limit of 8 iterations and bounds of 0 and ∞ .

2.4.3 Relationship with other results

If $V_\eta = 0$ then the results of Ref. [83] are regained. If $Q_i = 0$ and $L_i(\theta_i, \eta_i) = I_{N_c(\eta_i)}$ (I_n is the $n \times n$ identity matrix) then the standard results [8, 42, 53] on ML parameter estimation for Gaussian mixture problems by EM are regained.

2.5 Algorithm

The generalized EM algorithm operates by updating either (q, \bar{c}) , (q, V) , or (q, Q) . Because of the historical focus on \bar{c} in biology, the particular pattern of updating that is used focuses on \bar{c} . In particular, updating of (q, \bar{c}) is performed until \bar{c} converges in the sense of small quadratic norm of the difference between the current and the immediately previous value of \bar{c} . Then (q, V) and (q, Q) are updated. Then (q, \bar{c}) is updated. If the change in \bar{c} relative to its immediately previous value is sufficiently small to satisfy the convergence criteria then the algorithm terminates. Otherwise, the algorithm continues to update (q, \bar{c}) until convergence followed by another update of (q, V) and (q, Q) , *etc.* Joint optimization of Q and V was tested but works poorly possibly due to the difference in the sizes of these two covariance matrices. Pseudocode is given in Algorithm 1.

2.6 Performance

The standard measure of performance in structural biology is the Fourier Shell Correlation (FSC) [78, Eq. 2] [23, Eq. 17] [4, p. 879] between a pair of reconstructions computed from disjoint sets of images. The FSC is the spherical average of the correlation in the frequency domain between the two reconstructions normalized by the square root of the product of the spherical average of the magnitude squared of the individual reconstructions. The resolution of the reconstruction is defined to be the magnitude of the spatial frequency vector at the first crossing of a threshold where the threshold is typically taken to be $1/2$.

Algorithm 1: The generalized EM algorithm

set the initial condition from a homogeneous (i.e., $V = 0$) calculation

while true do

while \bar{c} not converged **do**

 update q (Eq. 2.26) and \bar{c} (Eq. 2.33)

end while

 update q (Eq. 2.26) and Q (Section 2.4.2)

 update q (Eq. 2.26) and V (Newton's method for $V_\eta = \text{diag}(v_\eta)$ in Section 2.4.2)

 update q (Eq. 2.26) and \bar{c} (Eq. 2.33)

if \bar{c} converged **then**

 break

end if

end while

Because existing reconstruction tools are based on having homogeneous objects (i.e., $V_\eta = 0$), the natural way in which to apply FSC to the heterogeneous (i.e., $V_\eta \neq 0$) work described in this chapter is to measure performance in terms of the mean, i.e., $\hat{\rho}_\eta(\mathbf{x})$.

Since this chapter concerns maximum likelihood (ML) estimators, the standard theory for the performance of ML estimators [18] can be applied. Let y be the vector of data and ω be the vector of unknown parameters. Let the estimate of ω , which is a function of y , be denoted by $\hat{\omega}(y)$. Let the Hessian of the log likelihood function, the matrix of mixed second-order partial derivatives of the log likelihood function, be denoted by $H(\omega)$ with i, j th element defined by $\partial^2 \ln p(y|\omega) / \partial \omega_i \partial \omega_j$ where $p(y|\omega)$ is the conditional probability density func-

tion on the data y given the unknown parameters ω . Let ω_* be the true value of the parameters. The key result [18] is that the estimation error, $\hat{\omega}(y) - \omega_*$, is approximately Gaussian distributed with mean vector $\mathbf{0}$ and covariance matrix $-[H(\hat{\omega}(y))]^{-1}$. Analogous to Ref. [51, Eq. 70], formulas for the Hessian can be derived so the covariance of the errors in \bar{c} can be determined. Furthermore, analogous to Ref. [51, Sections 4.3–4.4], a FSC-like criteria appropriate to ML can be derived based on the errors in \bar{c} .

Similarly, the Hessian with respect to V_η can be derived, and Eq. 2.44 is a special case for diagonal V_η . The matrix V_η must be positive semi definite. The most complicated V_η that is considered in Section 2.7 is diagonal so that the positive semi definite constraint simplifies to the constraint of non-negative diagonal elements. In the general case it may be more appropriate to both solve the ML optimization problem and analyze the estimation errors in terms of the Cholesky factor of V_η , for which the only constraint is the triangular structure. If $y = f(x)$ and the ML estimate of x is \hat{x} then the ML estimate of y is $f(\hat{x})$ [12, Theorem 7.2.10, p. 320]. Therefore, if \hat{C}_η is the ML estimate of the Cholesky factor then the ML estimate of V_η is $\hat{V}_\eta = \hat{C}_\eta \hat{C}_\eta^T$. Similarly, if \hat{V}_η is the ML estimate then $(\hat{V}_\eta)^{1/2}$ (where $^{1/2}$ indicates any algorithm for extracting the Cholesky factor) is the ML estimate of the Cholesky factor. Via the Cholesky factor, both the estimate of V_η and the analysis of the performance of the estimator will obey the positive-definite constraint on V_η . In the structural biology community, there is presently no consensus on a quantity that will summarize the estimator performance on the covariance analogous to the FSC summary of the estimator performance on the mean simply because that information has not previously been available.

2.7 Numerical results

Flock House Virus (FHV) [20,33,76] is an insect virus, hence a eukaryote virus, that has been intensively studied including both single-particle cryo electron microscopy [33] and x-ray crystallography [20] structures of the entire particle. Using the images and basis functions $(\{\phi_{\tau}^{(n)}(\mathbf{x})\}_{i=0}^{N-1}$ in Eq. 2.1) of Ref. [83, Section 3.1], in this section we describe a heterogeneous reconstruction of FHV. In summary of the calculation in Ref. [83, Section 3.1], there are 583 images each of 91×91 pixels with a sampling interval of 4.7Angstrom, the contrast transfer function is 1, there are 720 orthonormal basis functions described in spherical coordinates where the radial dependence is a spherical Bessel function and the angular dependence is an icosahedral harmonic and the basis function is supported in a sphere of radius 197.4Angstrom, there is one class, the nuisance parameters in θ are only the projection orientation, and the resulting resolution by FSC with threshold 1/2 is 27Angstrom. Other reconstructions described in Ref. [83, Section 3.1] achieve as high as 23Angstrom resolution from the same experimental images. For other examples, such as multi-class examples, please see Refs. [81,86]. Examples of the experimental images are shown in Figure 2.1(a).

A $V = 0$ calculation is performed in which the new software [a Matlab [40] program optimized to use primarily matrix-matrix operations running on a dual core dual cpu PC using 4 threads (4 "labs" in Matlab terminology)] yields the same structure as the software of Ref. [83]. Then the $V = 0$ constraint is removed.

Because viruses interact with other biological systems at the surface of the virus, the surface of FHV is important. Figure 2.1(b) shows the surface determined by the mean $\hat{\rho}_{\eta_0}(\mathbf{x})$ using pseudo-color determined by the covariance

$\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x})$. Therefore, Figure 2.1(b) includes a description of the heterogeneity of FHV that is a continuous heterogeneity within a single class of particle. On the surface of the particle (Figure 2.1(b)), the variance is highest near the 5-fold symmetry axes which is consistent with the idea that the binding of the particle to a new host cell and possibly later events concerning RNA translocation occur around a 5-fold axis [10, 13, 26]. Internal to the particle (Figure 2.1(d)), the variance is low on the ordered dodecahedral RNA cage [20, 68] that can be visualized in the x-ray crystallographic structure [20, 68], as expected, and higher in the unordered core of the particle.

The estimate of the mean, $\hat{\rho}_{\eta=1}(\mathbf{x})$, has icosahedral symmetry but this is not obvious in the sectioned visualization shown in Figure 2.1(c). Therefore, in Figure 2.2 we show 2-D cross sections through the origin normal to 5-, 3-, and 2-fold symmetry axes of the estimate $\hat{\rho}_{\eta=1}(\mathbf{x})$. In these images, the symmetry is clearly displayed.

The dynamic range of the 720 weights in $\bar{c}_{\eta=1}$ is great. Only one weight, the largest weight with value 12.13, has absolute value in the interval [10, 100). As the absolute value of the weights gets smaller, the number of weights gets larger and then smaller. Finally, there are 32 weights with absolute values in the interval [0, .01). The weights corresponding to spherically-symmetric low-spatial-frequency basis functions tend to be larger, corresponding to the fact that FHV is roughly spherical, which makes this a challenging reconstruction problem since it is challenging to detect the projection orientation direction for a roughly spherical object. The square roots of $v_{\eta=1}$ in $V_{\eta=1} = \text{diag}(v_{\eta=1})$ are neither constant in size nor a constant fraction of the size of the corresponding element in $\bar{c}_{\eta=1}$. Using the same intervals used to stratify the absolute values of the

weights, the intervals and the medians of the corresponding square roots (i.e., standard deviations) are [10, 100), 0.2387; [1, 10), 0.2581; [.1, 1), 0.3015; [.01, .1), 0.3082; and [0, .01), 0.3153. Thus the smaller coefficients, which are also the coefficients multiplying basis functions with higher spatial frequency content, tend to have larger variances. Note, however, that the covariance $V_{\eta=1}$ of the weights c is only indirectly related to the covariance Σ_i of the i th image y_i through the linear operator L_i (e.g., Eq. 2.7). Therefore it is difficult to make statements about the image based on knowledge of the covariance of the weights.

The value of λ (the pixel noise variance in $Q_{i,j} = \lambda I_{N_y}$) is changed less than 20% from its initial condition computed from the sample variance of the images [83, Section 2.8].

Allowing $V \neq 0$ could potentially change \bar{c} relative to a reconstruction with $V = 0$. However, in the case of FHV, this does not seem to be the case. In particular, Figure 2.3 shows the FSC curve comparing the structure mean structures (i.e., $\hat{\rho}_{\eta=1}(\mathbf{x})$) with $V = 0$ and $V \neq 0$ computed via Ref. [83, Eq. 25]. Continuing to use the threshold of 1/2, the conclusion is that the two structures are biologically equivalent.

2.8 Estimation of the *a priori* probability distribution on the nuisance parameters

Some objects, especially objects such as tailed bacteriophage that are far from spherical in shape, adopt a particular range of orientations relative to the air-water interface during the preparation of the cryo electron microscopy speci-

men. Therefore, the *a priori* probability distribution on the projection orientation nuisance parameters in θ_i is not uniform over all rotations (i.e., not Haar measure on $SO(3)$). In this section we generalize the class nuisance parameter η_i to allow estimation of the probability distribution of the projection orientation nuisance parameters.

Let H (capital η) be a nuisance parameter which has a many-to-one mapping to η (denoted by $\eta = \eta(H)$). The different values of H mapping to the same value of η are present so that H can separately select pdfs for θ , i.e., to make the pdf on θ into a mixture pdf. Specifically, each term in the mixture has the form $p(\theta, H) = p(\theta|H)p(H)$. However, both \bar{c} and V are functions of $\eta(H)$ not of H since it would be difficult to estimate large numbers of \bar{c} vectors and V matrices. The derivation of Section 2.4 can be repeated. An important change is that Eq. 2.22 is replaced by

$$p(\theta_i, H_i|y_i, \bar{c}, V, q, Q) = \frac{p(y_i|\theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)p(\theta_i|H_i)q_{H_i}}{\sum_{H'=1}^{N_H} \int_{\theta'} p(\theta'|H')q_{H'}p(y_i|\eta', \theta', \bar{c}^{\eta'}, V_{\eta'})d\theta'} \quad (2.47)$$

which implies that Eq. 2.22 becomes

$$p(\theta_i, \eta_i|y_i, \bar{c}, V, q, Q) = \sum_{\{H_i:\eta(H_i)=\eta_i\}} p(\theta_i, H_i|y_i, \bar{c}, V, q, Q). \quad (2.48)$$

Eq. 2.26 is replaced by

$$q_{H'} = \frac{1}{N_v} \sum_{i=1}^{N_v} \int_{\theta_i} p(\theta_i, H'|y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \alpha) d\theta_i \quad \forall H' \in \{1, \dots, N_H\} \quad (2.49)$$

where the computation of $p(\theta_i, H'|y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \alpha)$ is from Eq. 2.47. Eq. 2.33 (new estimate of \bar{c}^{η}), Eqs. 2.39–2.41 (Eqs. 2.43–2.46) (new estimate of V_{η}), and Eq. 2.30 (new estimate of Q) are unchanged but now the computation of $p(\theta_i, \eta'|y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \alpha)$ is from Eq. 2.48 rather than Eq. 2.22.

An important case is where each term in the mixture is impulsive. Since θ is a continuous random variable, this is an unusual choice. However, it com-

bines easily with the quadrature rules used to approximate integrals over θ . If all terms are impulsive, then without loss of generality, we can regard θ as discrete and perform all calculations in terms of Kronecker rather than Dirac delta functions. Making the further assumption that the same set of discrete values occurs for every class $\eta(H)$, it is natural to describe H by $H = (\eta, \xi)$ where η is the nuisance parameter of previous sections of the chapter and ξ selects the term in the mixture pdf for θ so that $p(\theta|H = (\eta, \xi)) = \delta_{\theta, \theta_\xi}$ where θ_ξ is the value of θ corresponding to mixture term index ξ . Since all terms are impulsive, the index of the term and the value of θ are equivalent so there exists a function $\xi(\theta)$. In this case, Eqs. 2.47 and 2.49 become

$$p(\theta_i, H_i = (\eta_i, \xi_i)|y_i, \bar{c}, V, q, Q) = \frac{p(y_i|\theta_{\xi_i}, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)\delta_{\theta_i, \theta_{\xi_i}}q_{H_i=(\eta_i, \xi_i)}}{\sum_{\eta'=1}^{N_\eta} \sum_{\xi'=1}^{N_\xi} q_{H'=(\eta', \xi')}p(y_i|\eta', \theta_{\xi'}, \bar{c}^{\eta'}, V_{\eta'})} \quad (2.50)$$

$$p(\theta_i, \eta_i|y_i, \bar{c}, V, q, Q) = \frac{p(y_i|\theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)q_{H_i=(\eta_i, \xi(\theta_i))}}{\sum_{\eta'=1}^{N_\eta} \sum_{\xi'=1}^{N_\xi} q_{H'=(\eta', \xi')}p(y_i|\eta', \theta_{\xi'}, \bar{c}^{\eta'}, V_{\eta'})}. \quad (2.51)$$

2.9 Discussion

In this chapter the first statistical model, estimator, and algorithm for computing reconstructions of *heterogeneous* biological objects, where the heterogeneity includes both discrete and continuous components, from single-particle cryo electron microscopy images is described and demonstrated. The problem formulation is equivalent to a Gaussian mixture parameter estimation problem where both the mean and covariance of each class are stochastically structured and the structuring involves complicated operations such as projection from 3 to 2 dimensions. A maximum likelihood criterion is used and the estimate is computed by an generalized expectation maximization algorithm. At least some biological applications of this approach will require computing on a large

scale, e.g., resolving heterogeneity at the 4Angstrom spatial scale in an object with a characteristic dimension of 500Angstrom based on 10^5 – 10^6 images each containing $10^2 \times 10^2$ pixels. Algorithmic and software engineering tradeoffs to enable such computations on a parallel cluster computer, e.g., choice of basis functions, Krylov versus direct solution of linear systems, appear to be fruitful areas for further investigation. Methods for characterizing the estimator performance also appear to be fruitful areas for further investigation.

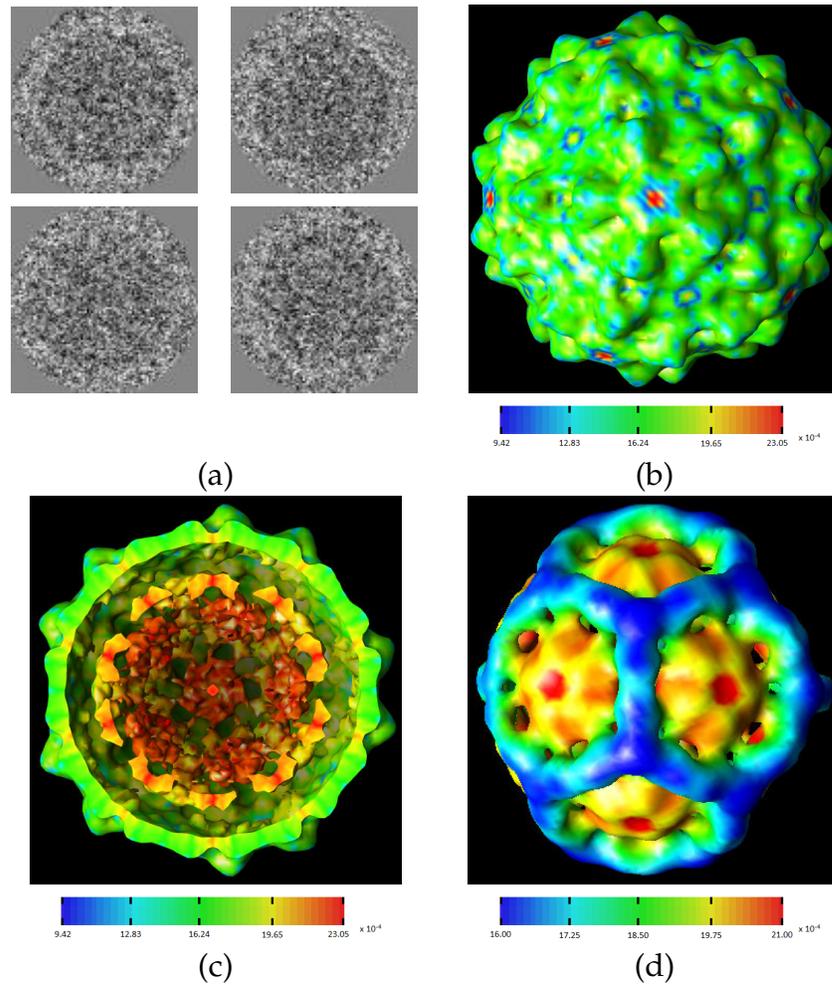
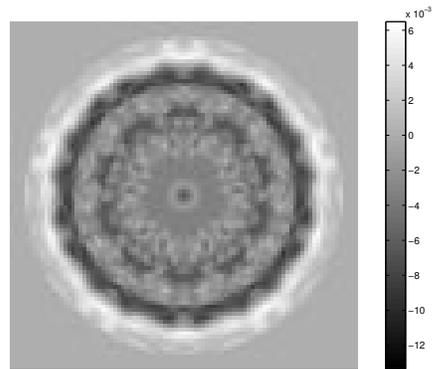
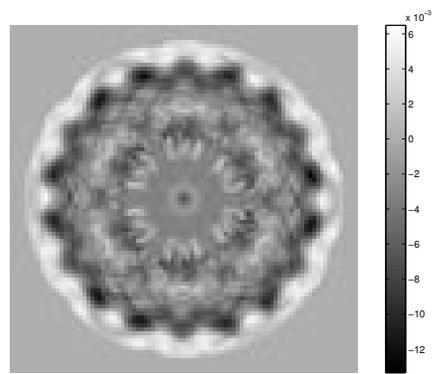


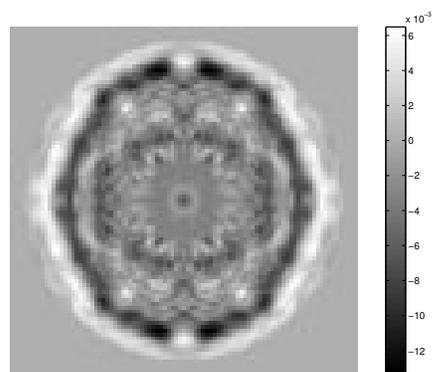
Figure 2.1: Reconstruction of FHV from experimental images. Panel (a): Example boxed experimental images (same color map). Panel (b): Surface plot of the mean $\hat{\rho}_{\eta_0}(\mathbf{x})$ pseudo-colored by the variance $\hat{r}_{\eta_0}(\mathbf{x}, \mathbf{x})$. The variance is highest near the 5-fold symmetry axes which is consistent with the idea that the binding of the particle to a new host cell and possibly later events concerning RNA translocation occur around a 5-fold axis [10, 13, 26]. Panel (c): Cross section of the mean though the center of the particle perpendicular to a 5-fold symmetry axis pseudo-colored by the variance. Panel (d): The RNA core after removal of the protein capsid. Surface plot of the mean pseudo-colored by the variance. The ordered dodecahedral RNA cage [20, 68] is detected and, as expected, the variance of the cage is low (blue) while that of the surrounding less well-ordered material is high (red). Visualizations in Panels (b–d) by UCSF Chimera [49].



5-fold



3-fold



2-fold

Figure 2.2: Cross sections through the origin normal to 5-, 3-, and 2-fold symmetry axes of the estimated 3-D mean ($\hat{\rho}_{\eta=1}(\mathbf{x})$) function. The symmetries can clearly be seen in the cross sections, although the 5-fold symmetry is approximately a 10-fold symmetry and the 3-fold symmetry is approximately a 6-fold symmetry. The same color map is used in all images.

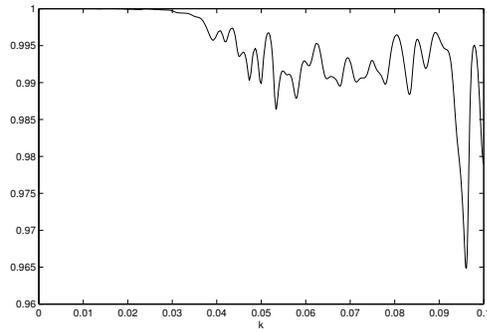


Figure 2.3: Fourier Shell Correlation (FSC) curve between $\hat{\rho}_{\eta=1}(\mathbf{x})$ computed with V constrained to $V = 0$ versus the optimal diagonal V . The independent variable k is the magnitude of the spatial frequency vector measured in Angstrom^{-1} . The pixel size of 4.7Angstrom implies a Nyquist frequency of $1/(2 \times 4.7) = 0.106\text{Angstrom}^{-1}$ which determined the upper limit of $k = 0.1\text{Angstrom}^{-1}$. Since the curve remains above $1/2$ for the entire range $k \in [0, 0.1]\text{Angstrom}^{-1}$, from the biological point of view the two mean structures are equivalent.

CHAPTER 3

APPLICATIONS ON HETEROGENEOUS VIRUS PARTICLES DURING THEIR MATURATION OVER TIME

3.1 Heterogeneous *Nudaurelia Capensis* ω Virus ($N\omega V$)

3.1.1 Introduction

Recent success with 3-dimensional reconstructions of biological macro molecular particles employing single-particle cryo electron microscopy (cryo EM) has been remarkable. Sub nanometer icosahedral virus structures are virtually routine and protein and nucleo-protein structures, without symmetry, are appearing more frequently at comparable resolution. Structures of icosahedral viruses at near-atomic resolution have been achieved with this technology in recent years [3,39,84].

Cryo EM data captures biological macromolecular particles that are trapped in one of a smooth continuum of conformations at the moment of vitrification in liquid ethane. The amount of conformational change accessible to the particle is presumably space dependent, but there are limited tools available for assessing the global amount of conformational change available let alone creating a spatial map of the amount of conformational change occurring.

Cryo EM has recently been reviewed in the three volumes edited by Jensen [28–30]. The idea of maximum likelihood as a method for deriving statistical estimators dates back to the early 1900s [37, Section 10.1, p. 515] and it remains an important method. Computation of a reconstruction by optimization

of the fit between the images predicted by a mathematical model and the experimental images, which can be interpreted as a maximum likelihood estimator, was first done by Vogel, Provencher, Bonsdorff, Adrian, and Dubochet [79] and Provencher and Vogel [50, 80] and has recently been reviewed [55, 59]. Maximum likelihood has also been used for other estimation tasks related to cryo electron microscopy, such as estimating the orientation of an image [58]. Heterogeneity among a set of particles can be detected by methods such as cross-common lines residuals [22]. In this chapter, maximum likelihood estimation is used not to estimate a single reconstruction or to find a homogeneous subset of particles but rather to estimate the statistics of an entire ensemble of reconstructions where the statistics of the images predicted by the statistics of the ensemble of reconstructions match the statistics of the experimental images. The most closely related work is due to Penczek, Yang, Frank, and Spahn [48]. In this work, a space-varying variance map was constructed after the reconstruction is computed by a Monte-Carlo resampling procedure. This contrasts with the approach proposed in this thesis where the mean and covariance information are simultaneously estimated, generating not only the reconstruction but also the variance associated with every voxel of the reconstruction.

The method was used to reanalyze the time-resolved single-particle cryo EM images of Nudaurelia Capensis Omega Virus ($N\omega V$) from Matsui, Lander, Khayat, and Johnson [41], a $T = 4$ icosahedral RNA virus. $N\omega V$ capsid is composed of 240 copies of the same gene product, protein alpha, that in a maturation step, undergoes a autocatalytic reaction generating the major capsid protein beta and the small gamma peptide, which remains non-covalently associated with the capsid. $N\omega V$ virus-like particles can be purified in the uncleaved pro-capsid state and the maturation process can be precisely triggered by low-

ering the pH to 5.0. Kinetics of the cleavage is unusual with 50% of the subunits cleaved in 30 minutes while several hours are required for all of the subunits to cleave. Taking advantage of the slow kinetics of maturation of N ω V, partially cleaved particles in intermediate stages of maturation (3 minutes, 30 minutes, and 4 hours all at pH 5.0) were analysed by cryo-EM. Because the size of the particles is the same throughout the maturation process, it was possible to use difference cryo-EM density maps. The density at each time point was subtracted from the fully mature particle. With the x-ray model as a guide, the difference density at each of the cleavage sites was evaluated. Subunits surrounding 5-fold and 3-fold icosahedral symmetry axes are quickly formed and cleave in 30 minutes, while the subunits not adjacent to these axes cleave slowly. Here, we show that the maximum-likelihood derived variance map can, in a single data set, reveal the same local variations that were observed with difference map analysis, and also provide an overall view of particle dynamics that was unobservable with classical methods of analysis. The data are the time-resolved single-particle cryo EM images of N ω V from Matsui, Lander, Khayat, and Johnson [41]. The pixels measure 2.768Angstrom and the boxed image of an individual particle is 200 \times 200 pixels in dimension. The reconstructions from Matsui *et al.* (2010) have been deposited in the EM Data Bank [70]. The times, reference numbers, and accession codes are 3 min, 25633, EMD-5426; 30 min, 25634, EMD-5427; 4 h, 25635, EMD-5428; and 3 d, 25622, EMD-5425, respectively. The assumption that there is only one class of particle at each time point was sufficient to achieve resolutions of 9.3Angstrom, 8.6Angstrom, 8.3Angstrom, and 9.8Angstrom for the 3 minutes, 30 minutes, 4 hours, and 3 days data sets, respectively (Matsui, Lander, Khayat, and Johnson [41], and so the calculations described in this chapter, which are at lower resolution, have continued with that assumption. The re-

sults from the calculation described in this chapter have been deposited in the EM Data Bank [70]. Each data set results in two depositions: a mean map and a variance map. For the mean maps, the times, reference numbers, and accession codes are 3 min, 25729, EMD-5449; 30 min, 25858, EMD-5474; 4 h, 25859, EMD-5472; and 3 d, 25860, EMD-5473, respectively. For the variance maps, the times, reference numbers, and accession codes are 3 min, 25861, EMD-5468; 30 min, 25863, EMD-5469; 4 h, 25864, EMD-5471; and 3 d, 25865, EMD-5470, respectively.

3.1.2 Computational methodology

Using a weighted sum of basis functions to represent the electron scattering intensity function has a long history in structural biology, e.g., Fourier series in x-ray crystallography. If every instance of the object is identical, then the weights in the description of each object are the same and the goal of structure determination is to determine the numerical value of each weight. But if different instances of the object are different, then there is no unique numerical value for each weight. Different instances might differ by different stoichiometry or by different geometrical configuration, e.g., flash frozen in different vibrational conditions for single-particle cryo EM problems. If the differences can be described as statistical variation, then the goal of structure determination might be to determine the numerical values of the means and variances of each weight. If the weights are assumed to be Gaussian random variables and are grouped in a vector, then the mean vector and covariance matrix for the weight vector is a complete description of the object.

The change from describing the weights as numbers and estimating the numbers for each class of object, to describing the weights as Gaussian and estimating the statistics (the mean, corresponding to a traditional reconstruction, and the covariance, describing fluctuations around the reconstruction) for each class of object is the modeling innovation proposed in this thesis. Using this new model, a maximum likelihood estimator is used to determine the means and covariances that are the solution of the reconstruction problem and the estimator is computed by a generalized expectation maximization algorithm which is an iterative algorithm which must be provided with an initial condition. The pixel noise variance and the probability that an image belongs to a particular class are also estimated. Optionally, but not used in the calculations described in this chapter, the *a priori* probability density function on the projection orientation of the images can also be estimated. In the expectation maximization algorithm, simultaneous updates of all parameters to be estimated is a difficult optimization problem so the mean vector, the covariance matrix, and the pixel noise variance are updated sequentially (so that this is actually a generalized expectation maximization algorithm). Each update is the solution of a maximization problem. For the mean vector, the maximization problem is quadratic in the unknown vector so the new mean vector is the solution of a linear system quite similar to the situation in the homogeneous particle case [16,34,36,51,83]. For the covariance matrix, the maximization problem is complicated because the covariance of an image is a linear combination of the covariance of the weights in the orthonormal expansion and the variance of the additive pixel noise. The linear combination is unknown and is different for each different image. Intuitively, the observed variability in the images is being partitioned into two sources which are the heterogeneity of the particle and the additive pixel noise.

Because the covariance, rather than the inverse covariance, is a linear combination of the covariance of the weights in the orthonormal expansion and the variance of the additive pixel noise, this maximization problem is not convex. Formulas for first and second derivatives of the function to be maximized with respect to the covariance of the weights can be determined and, using the function and the derivatives, the maximization problem is solved numerically. Finally, for the additive pixel noise covariance, a search based on just the function to be maximized is used since the unknown is a scalar and an accurate initial condition is available by computing the sample variance of the pixels in the images in an annulus outside of the image of the particle. A flow chart of the algorithm is given in Figure 3.1. Figure 3.1(a) shows the entire algorithm with preprocessing, repetition of reconstruction calculations on non-overlapping sets of boxed images in order to provide the data necessary for computing sample variances, and postprocessing. Figure 3.1(b) shows the reconstruction algorithm for a single set of boxed images.

In order to compute the performance of the algorithm, at each time point the algorithm is run on each of four distinct data sets where the data sets are nonoverlapping subsets of the four image stacks of Matsui, Lander, Khayat, and Johnson [41]. Then, based on the four results, sample standard deviations can be computed which describe the performance of the algorithm. This overall computation is shown in Figure 3.1(a). The algorithm is iterative and therefore requires an initial condition. At each time point, for the first of the four data sets, the algorithm is used twice: (1) The algorithm is started with means that describe a spherically-symmetric reconstruction and zero covariances and is run to the final resolution with the heterogeneity features turned off. In this case the algorithm is equivalent to the algorithm of Refs. [16,83]. Alternatively, this cal-

ulation could be described as Block 2 of the flow chart in Figure 3.1(b) or the third line of Algorithm 2 with the addition of the standard idea of refinement where the resolution of the reconstruction is progressively increased. (2) The algorithm is restarted with the means equal to the solution from Step (1) and the variances equal to 10% of the corresponding means and is run with the heterogeneity features turned on to determine the heterogeneous reconstruction. At each time point, for the second through fourth data sets, only Step (2) is used starting from the homogeneous reconstruction resulting from Step (1) applied to the first data set since there is no need to find a new initial condition.

Once the mean vector and covariance matrix for the weights in the orthonormal expansion have been estimated, the nominal structure can be computed from the mean vector and the variance map can be computed from the covariance matrix.

The following subsections describe the computational methods in detail, including pre- and post-processing that go beyond the ideas of Chapter 2.

Preprocessing

Subsequent to the steps used to create the image stacks in Matsui, Lander, Khayat, and Johnson [41], the following procedure was carried out.

1. Reconstruction algorithms can include provisions for rejecting images from the image stack because the images appear not to belong to the particle as it appears in the 3-D reconstruction being computed. Such provisions are not included in the current version of the reconstruction algorithm described in this chapter. Therefore, some possibly junk images are

removed from the stack before the reconstruction algorithm begins by the following mechanism: First, select the first 6000 images from the stack. Second, compute the sample mean of all the selected images. (The mean image is nearly circularly symmetric). Third, compute the difference between each particular image and the sample mean image. Fourth, compute the square of the Euclidean norms of the difference images. Fifth, from a histogram of the squared norms, decide on a threshold and remove images from the stack if their squared norm is greater than the threshold. About 16% of the images are removed. For the stack recorded at 3 minutes, 20 of the removed images are shown in Figure 3.12.

2. In order to compute the performance of the algorithm, e.g., the error bars of Figure 3.6, the algorithm is applied to multiple sets of images. Specifically, from those images that are not removed from the stack, we form four substacks each with 1200 images by first randomly permuting the 6000 images and then selecting subsets of 1200 images where the subsets are those images numbered $4n - 3$, $4n - 2$, $4n - 1$, and $4n$ where $n \in \{1, \dots, 1200\}$.
3. Individually for each image in a stack, normalize the image. Specifically, $y^{\text{new}} = ay^{\text{old}} + b$ where a and b are chosen so that the sample mean and the sample variance of y^{new} , both evaluated outside of the image of the virus particle, have values 0 and 1, respectively.

Reconstruction

In this section, the specific version of the methods of Chapter 2 that was applied to the $N\omega V$ problem is described.

The electron scattering intensity of the particle is described as a weighted

sum of basis functions. A standard approach is to treat the set of weights as numbers and seek to estimate the values of the numbers by a maximum likelihood estimator [16, 34, 36, 51, 57, 83]. In contrast, in this thesis we treat the set of weights as random variables where every particle is described by an independent realization of the random variables. We then seek to estimate the joint probability density function of the set of random variables. In order to simplify the task from estimating functions to estimating numbers, we assume that the joint probability density function is Gaussian so that all we must estimate is the mean vector and covariance matrix. These quantities can be estimated by a maximum likelihood estimator which is computed by an expectation-maximization algorithm where the nuisance parameters in the expectation-maximization algorithm include the unknown projection direction of the image of each particle.

Reconstruction: Notation If x is a random variable then $x \sim p$ means that x is distributed with probability density function (pdf) p . $\mathcal{N}(m, S)(x)$ is the Gaussian pdf with mean vector m and covariance matrix S evaluated at argument x . If v is a vector (which might already have multiple superscripts and subscripts) then $(v)_j$ is the j th component of the vector. Likewise, if M is a matrix then $(M)_{j,j'}$ is the (j, j') th element of the matrix.

Reconstruction: Model The electron scattering intensity ρ as a function of 3-D real-space coordinates \mathbf{x} is described by a truncated orthonormal expansion with weights c and basis functions ϕ :

$$\rho^{(\eta)}(\mathbf{x}) = \sum_{j=1}^{N_c(\eta)} c_j^{(\eta)} \phi_j^{(\eta)}(\mathbf{x}) \quad (3.1)$$

where η is the class label and there are N_η classes. In first-order image formation theory [19, 38, 69], the reciprocal-space image, denoted by Υ , parameterized by the 2-D reciprocal-space vector, denoted by $\boldsymbol{\kappa}$, is the product of three factors. (1) The 2-D Fourier transform of the projection image which, by the projection slice theorem, can be computed from the 3-D Fourier transform P of the object ρ and the 3×3 rotation matrix R that describes the projection direction which is parameterized by the Euler angles (α, β, γ) . (2) The contrast transfer function G . (3) A complex exponential of the translation $\boldsymbol{\chi}_0$ of the projected location of the center of the object from the center of the reciprocal space image. The resulting equation is

$$\Upsilon_i(\boldsymbol{\kappa}) = \exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(|\boldsymbol{\kappa}|)P^{(\eta_i)}\left(R_{\alpha_i,\beta_i,\gamma_i}^{-1}\begin{bmatrix}\boldsymbol{\kappa} \\ 0\end{bmatrix}\right). \quad (3.2)$$

In order to make Eqs. 3.1 and 3.2 into numerical linear algebra, the spatial frequency vector $\boldsymbol{\kappa}$ is discretized and Eq. 3.2 for each sample is one row of the resulting vector equation. In addition, the notation is augmented with an index i which indicates which of the boxed images is being described and Φ is the 3-D Fourier transform of the basis function ϕ . The resulting equation is

$$y_i = L(z_i)c^{(\eta_i)} \quad (3.3)$$

where

1. y_i is a vector whose j th component is the the reciprocal space image evaluated at the j th sampled reciprocal space vector $\boldsymbol{\kappa}_j$, i.e.,

$$(y_i)_j = \Upsilon_i(\boldsymbol{\kappa}_j). \quad (3.4)$$

2. z_i is the Euler angles $(\alpha_i, \beta_i, \gamma_i)$ that describe the projection orientation of the i th image, the 2-component vector $\boldsymbol{\chi}_{0,i}$ that describes the projected location

of the center of the particle in the i th image, and the class label η_i , i.e.,

$$z_i = (\alpha_i, \beta_i, \gamma_i, \chi_{0,i}, \eta_i), \quad (3.5)$$

all of which are unknown.

3. $c^{(\eta_i)}$ is the vector of weights for the i th particle, i.e.,

$$(c^{(\eta_i)})_j = c_j^{(\eta_i)}. \quad (3.6)$$

4. $L(z_i)$ is the matrix that describes the transformation from weights to sampled reciprocal-space image as is given in Eq. 3.2, i.e., weights to 3-D cube, projection from 3-D to 2-D, the effect of the contrast transfer function, and the translation of the projected location of the center of the particle in the i th image so the (j, j') th element of this matrix is

$$(L(z_i))_{j,j'} = \exp(-i2\pi\boldsymbol{\kappa}_j^T \chi_{0,i}) G(|\boldsymbol{\kappa}_j|) \Phi_{j'}^{(\eta_i)} \left(R_{\alpha_i, \beta_i, \gamma_i}^{-1} \begin{bmatrix} \boldsymbol{\kappa}_j \\ 0 \end{bmatrix} \right). \quad (3.7)$$

The statistical model used previously [16,57]) is that every object in the η th class is identical and the projection image is corrupted by additive zero-mean Gaussian noise that is independent from image to image. The *a priori* probability that an object is from the η th class is q_η . The resulting equations are

$$y_i = L(z_i)c^{(\eta_i)} + v_i \quad (3.8)$$

$$v_i \sim \mathcal{N}(0, Q). \quad (3.9)$$

where the goal is to estimate the vectors $c^{(\eta)}$ for $\eta \in \{1, \dots, N_\eta\}$. This problem can be generalized to include estimating the *a priori* probability density function on the orientation of the projections and estimating the *a priori* probabilities of each class [57].

In this thesis it is proposed to allow each instance of an object in the η th class to have a different structure where the variability is described statistically by assuming that the weights for the orthonormal expansion (Eq. 3.1) collected into a vector (Eq. 3.6) are Gaussian random vectors with mean vector \bar{c}^η and a covariance matrix V_η . The resulting equations are

$$y_i = L(z_i)c_i + v_i \quad (3.10)$$

$$c_i \sim \mathcal{N}(\bar{c}^{\eta_i}, V_{\eta_i}) \quad (3.11)$$

$$v_i \sim \mathcal{N}(0, Q) \quad (3.12)$$

where the c_i random vectors are nuisance parameters, that is, they are not known but instead of estimating them, a pdf for them is provided. Since linear transformations of Gaussian random vectors are Gaussian random vectors, rewrite Eqs. 3.10–3.12 with a single Gaussian random vector v' rather than two Gaussian random vectors c and v . The resulting equations are

$$y_i = L(z_i)\bar{c}^{\eta_i} + v'_i \quad (3.13)$$

$$v'_i \sim \mathcal{N}(0, L(z_i)V_{\eta_i}L^T(z_i) + Q). \quad (3.14)$$

Eqs. 3.10–3.12 and Eqs. 3.13–3.14 differ in two important ways. First, the c_i random vectors are gone leading to simpler estimator equations. Second, v' has a structured covariance matrix, specifically, $L(z_i)V_{\eta_i}L^T(z_i) + Q$. The goal is to estimate Q , q_η , \bar{c}^η , V_η for $\eta \in \{1, \dots, N_\eta\}$. In addition, though it is not done in this chapter, it is possible to estimate the *a priori* pdf on the orientation of the projections.

Reconstruction: Estimator Using the notation of Section 3.1.2, it follows from Eqs. 3.13 and 3.14 that the conditional mean, denoted by $\mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i})$, and the

conditional covariance, denoted by $\Xi_i(\theta_i, \eta_i, V_{\eta_i}, Q_i)$, of the i th image, denoted by y_i , are

$$\mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}) \doteq E[y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i] \quad (3.15)$$

$$= L_i(\theta_i, \eta_i) \bar{c}^{\eta_i} \quad (3.16)$$

$$\Xi_i(\theta_i, \eta_i, V_{\eta_i}, Q_i) \doteq \text{Cov}[y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i] \quad (3.17)$$

$$= L_i(\theta_i, \eta_i) V_{\eta_i} L_i^T(\theta_i, \eta_i) + Q_i \quad (3.18)$$

where the operators E and Cov are expectation and covariance, respectively, and that the conditional probability density function (pdf) on y_i is

$$p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) = \mathcal{N}(\mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}), \Xi_i(\theta_i, \eta_i, V_{\eta_i}, Q_i))(y_i). \quad (3.19)$$

The absence of a subscript or superscript implies that the variable is the collection of variables with the subscript or superscript, e.g., $\bar{c} = (\bar{c}^{\eta}|_{\eta=1}, \dots, \bar{c}^{\eta}|_{\eta=N_{\eta}})$. In this abbreviated notation, the log likelihood function for the maximum likelihood estimator is

$$\ln p(y | \bar{c}, V, q, Q) = \sum_{i=1}^{N_v} \ln \left[\sum_{\eta_i=1}^{N_{\eta}} \int_{\theta_i} p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) q_{\eta_i} p(\theta_i) d\theta_i \right] \quad (3.20)$$

where N_v is the number of particles that are imaged, $p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i)$ is given in Eq. 3.19, and $p(\theta_i)$ is the *a priori* pdf on θ_i and the definition of the estimator is

$$\hat{c}, \hat{V}, \hat{q}, \hat{Q} = \arg \max_{\bar{c}, V, q, Q} \ln p(y | \bar{c}, V, q, Q) \quad (3.21)$$

where the $\hat{\cdot}$ indicates that the variable is an estimate.

The method used for computing the maximization is a generalized expectation-maximization algorithm. The idea in expectation-maximization algorithms is that there is a set of so-called nuisance parameters which, if their values were measured, would greatly simplify the computation of the maximum. However, the values are not measurable. The iterative nature of the algorithm results from repeating a pair of steps: average over the possible values

of the nuisance parameters (the so-called expectation step) and compute new values for the parameters being estimated by maximizing the result of the averaging with respect to the parameters. For this problem, the natural nuisance parameters are the variables θ_i, η_i ($i \in \{1, \dots, N_v\}$).

The conditional pdf on the nuisance parameters is

$$p(\theta_i, \eta_i | y_i, \bar{c}, V, q, Q) = \frac{p(y_i | \theta_i, \eta_i, \bar{c}^{\eta_i}, V_{\eta_i}, Q_i) p(\theta_i) q_{\eta_i}}{\sum_{\eta'=1}^{N_\eta} \int_{\theta'} q_{\eta'} p(\theta') p(y_i | \eta', \theta', \bar{c}^{\eta'}, V_{\eta'}, Q_i) d\theta'} \quad (3.22)$$

which uses Eq. 3.19. Using Eq. 3.22 repeatedly and following the calculation of Ref. [16], the update equations for the generalized expectation maximization algorithm are described in the following paragraphs. In all the following equations, variables with a leading subscript of 0, e.g., ${}_0V$, are the result of the previous iteration and variables without a leading subscript of 0, e.g., V , are the variables being computed in the current iteration.

(1) For each class (equivalently, each value of η' in the set $\{1, \dots, N_\eta\}$), the new value of the *a priori* class probability, denoted by $q_{\eta'}$ as a function of $\bar{c}, {}_0V, \mathcal{Q}$, and ${}_0Q$ is

$$q_{\eta'} = \frac{1}{N_v} \sum_{i=1}^{N_v} \int_{\theta_i} p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) d\theta_i \quad (3.23)$$

where the computation of $p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i)$ is from Eq. 3.22. The primary computational expense is to compute the integrals in Eqs. 3.23, 3.25, 3.26, and 2.35. (Figure 3.1, Blocks 2, 3, and 4).

(2) The new value of \bar{c} as a function of $V, Q, \bar{c}, {}_0V, {}_0Q$ is determined by solving the following linear system for each $\eta' \in \{1, \dots, N_\eta\}$ to compute the corresponding $\bar{c}^{\eta'}$ vectors:

$$F(\eta', y, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) \bar{c}^{\eta'} = g(\eta', y, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) \quad (3.24)$$

where

$$\begin{aligned}
& F(\eta', y, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) \\
&= \sum_{i=1}^{N_v} \int_{\theta_i} L_i^T(\theta_i, \eta') \Xi_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) L_i(\theta_i, \eta') p(\theta_i, \eta' | y_i, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) d\theta_i
\end{aligned} \tag{3.25}$$

$$\begin{aligned}
& g(\eta', y, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) \\
&= \sum_{i=1}^{N_v} \int_{\theta_i} L_i^T(\theta_i, \eta') \Xi_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) y_i p(\theta_i, \eta' | y_i, \bar{\sigma}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}, {}_0Q_i) d\theta_i.
\end{aligned} \tag{3.26}$$

(Figure 3.1, Block 2).

(3) The new value of V as a function of \bar{c} , Q , \bar{c} , ${}_0V$, ${}_0Q$ is computed by nonlinear programming. First define

$$N_i(y_i, \theta_i, \eta_i, \bar{c}^{\eta_i}) \doteq (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i})) (y_i - \mu_i(\theta_i, \eta_i, \bar{c}^{\eta_i}))^T \tag{3.27}$$

and

$$\begin{aligned}
& Q_1(\bar{c}, V, q, Q | \bar{c}, {}_0V, \mathcal{Q}, {}_0Q, y) \\
&= -\frac{N_y}{2} \ln(2\pi) N_v + \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_{\eta}} \ln \det(\Xi_i^{-1}(\theta_i, \eta_i, V_{\eta_i}, Q_i)) p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{Q}, {}_0Q_i) d\theta_i \\
&\quad - \frac{1}{2} \sum_{i=1}^{N_v} \int_{\theta_i} \sum_{\eta_i=1}^{N_{\eta}} \text{tr} \left[\Xi_i^{-1}(\theta_i, \eta_i, V_{\eta_i}, Q_i) N_i(y_i, \theta_i, \eta_i, \bar{c}^{\eta_i}) \right] p(\theta_i, \eta_i | y_i, \bar{c}^{\eta_i}, {}_0V_{\eta_i}, \mathcal{Q}, {}_0Q_i) d\theta_i.
\end{aligned} \tag{3.28}$$

Then the new value of V_{η} is the value that maximizes Eq. 3.28. (Figure 3.1, Block 4).

(4) The new value of Q as a function of V , \bar{c} , \bar{c} , ${}_0V$, ${}_0Q$ is only considered for the case where the pixel noise is independent and identically distributed at all pixels of all images. Then Q is just a scalar covariance which is denoted by λ and which must be determined by nonlinear programming to maximize the value of Eq. 3.28. (Figure 3.1, Block 3).

Reconstruction: Algorithm These four steps of Section 3.1.2 can be combined in many ways to yield valid expectation maximization algorithms. Focusing on the importance of the mean vector, which is the traditional reconstruction, the calculations described in this chapter use the algorithm described in Algorithm 2. Several aspects of Algorithm 2 need additional explanation. The

Algorithm 2: The generalized EM algorithm

set the initial conditions on \bar{c} , V , q , and Q .

while true do

while \bar{c} not converged **do**

 update q (Eq. 3.23) and \bar{c} (Eq. 3.24) (Figure 3.1, Block 2)

end while

while q and Q not converged **do**

 update q (Eq. 3.23) and Q . (Figure 3.1, Block 3)

end while

while q and V not converged **do**

 update q (Eq. 3.23) and V . (Figure 3.1, Block 4)

end while

 update q (Eq. 3.23) and \bar{c} (Eq. 3.24) (Figure 3.1, Block 2)

if \bar{c} converged **then**

 break

end if

end while

algorithm is an *ab initio* algorithm but it has not often been used in that mode. Instead, it has typically been used based on a traditional homogeneous reconstruction which provides a high-quality estimate of mean \bar{c}^n for each value of

$\eta \in \{1, \dots, N_\eta\}$ and this estimate is used as the \bar{c}^η initial condition. Because the optimization problem is for the covariance V_η and not, for instance, the Cholesky factor of V_η , it is necessary to impose the constraint that V_η be semi positive definite. Therefore, the biologically-natural initial condition of $V_\eta = 0$ is on the boundary of the feasible set and the nonlinear programming algorithms that have been used do not behave well in this situation. Therefore, the initial condition that has been used is a diagonal initial condition where the j th element is 10% of the j th element of the \bar{c}^η initial condition. The initial condition for Q , the pixel noise, is the sample variance in an annulus of the image surrounding the portion of the image that displays the virus particle, averaged over all the images in the calculation. The initial condition for q_η , the class probability, is uniform, i.e., $q_\eta = 1/N_\eta$ for each value of $\eta \in \{1, \dots, N_\eta\}$.

Reconstruction: Software The theory of earlier subsections applies for any choice of basis functions. However, the software uses the specific basis functions described in Ref. [83] where each basis function is the product of an icosahedral harmonic and a spherical Bessel function. A software implementation of the method was written that is suitable for execution in either the proprietary Matlab [40] or open source Octave [47] engine on a shared-memory computer. The update of Q is done by `fminbnd` in both Matlab and Octave. The update of V is implemented only for the case where V is a diagonal matrix and is done by `fmincon` in Matlab and `SQP` in Octave. The limits of the software are partly memory requirements and partly use of simple numerical linear algebra algorithms. As an example of algorithmic limitations, Eq. 3.24 is solved by LU decomposition where the F matrix (Eq. 3.25) and g vector (Eq. 3.26) are each computed without taking advantage of the fact that orientations that are close to

each other lead to similar contributions to the integrals in Eqs. 3.25 and 3.26. Relative to memory requirements, in order to make efficient Matlab code, the data is treated as matrix with dimensions that are the number of pixels per image by the number of images. This is the largest data structure in the runs described in this chapter. With the exception of the results for HK97 which are described in Section 3.2.2, all of the results described in this thesis are based on running the software using the Matlab engine on a dual-cpu quad-core Xeon (E5430 at 2.66GHz) with 16GB memory. In order to fit a computation into this hardware-software system, using more images implies using fewer basis functions or visa versa. For the results described in this chapter, all calculations used 1200 images and 720 basis functions (the so-called Step 7 of Ref. [83]) and each reconstruction takes approximately 2 days.

Reconstruction: Comments In the work of Penczek, Yang, Frank, and Spahn [48], a space-varying variance map is constructed by a Monte-Carlo resampling procedure after the reconstruction is computed while in the approach proposed in this thesis, the mean and covariance are simultaneously estimated. Potentially, though not demonstrated in the example of Section 3.1.3, the simultaneous estimation will allow for a better reconstruction since the reconstruction algorithm is allowed the additional degrees of freedom of assigning high variance to a part of the structure rather than allowing the somewhat disordered state of a part of the structure to contaminate better ordered parts of the structure. A second contrast is that the information estimated in the calculations of this chapter is sufficient to construct the complete second-order statistics of the reconstruction, i.e., a space-varying mean (the reconstruction) and a space-varying autocorrelation function. The autocorrelation function is the covariance

between the electron scattering intensity at two different locations and therefore is a function of 6 independent variables (two 3-D spatial positions). It would be very challenging to estimate such a large amount of information by resampling. An advantage of resampling is that very little must be assumed about the probability density functions. However, the assumptions that are made in this thesis have a long history (dating back to at least 1984 [53]) in the pattern recognition and machine learning communities as assumptions that are still useful even if there is no underlying physical model to motivate them.

The Gaussian assumption used in the homogeneous case [16, 34, 36, 51, 57, 83] (Eq. 3.9) greatly simplifies the maximization step of the expectation-maximization algorithms but may not have a more fundamental motivation. Here, however, the joint Gaussian assumption on the pixel noise and weights (Eqs. 3.11 and 3.12) is important because it allows the combination of these two sources of variability into a single equivalent source (Eq. 3.14).

For the reconstruction of homogeneous particles (Eqs. 3.8 and 3.9), a fast algorithm exists [34] that takes advantage of the fact that one of the Euler angles corresponds to a rotation of the image in the plane of the image. However, no corresponding algorithm appears to be possible for reconstruction of heterogeneous particles (Eqs. 3.13 and 3.14).

Postprocessing

In order to interpret the results, estimates of the statistics of the weights in the orthonormal expansion are not as intuitive as estimates of the statistics of the electron scattering intensity function. Conditional on a particular class, the spa-

tial mean function (which depends on position in 3-D space) and the spatial variance function (which depends on position in 3-D space) of the electron scattering intensity are

$$\bar{\rho}_{\eta'}(\mathbf{x}) \doteq \text{E}[\rho(\mathbf{x})|\eta = \eta'] \quad (3.29)$$

$$= \sum_{j=1}^{N_c(\eta')} (\bar{c}^{\eta'})_j \phi_j^{(\eta')}(\mathbf{x}) \quad (3.30)$$

and

$$v_{\eta'}(\mathbf{x}) \doteq \text{E}[[\rho(\mathbf{x}) - \bar{\rho}_{\eta'}(\mathbf{x})]^2|\eta = \eta'] \quad (3.31)$$

$$= \sum_{j=1}^{N_c(\eta')} \sum_{j'=1}^{N_c(\eta')} (V_{\eta'})_{j,j'} \phi_j^{(\eta')}(\mathbf{x}) \phi_{j'}^{(\eta')}(\mathbf{x}), \quad (3.32)$$

respectively. The variance function is a special case of the correlation function of Eq. 3.32. Let $\hat{\rho}_{\eta'}(\mathbf{x})$ and $\hat{v}_{\eta'}(\mathbf{x})$ be Eqs. 3.30 and 3.32 evaluated at the estimated values of \bar{c} and V rather than the true values. For biological purposes, the natural quantities to visualize are $\hat{\rho}_{\eta'}(\mathbf{x})$ and $\hat{v}_{\eta'}(\mathbf{x})$, especially the standard deviation $s_{\eta'}(\mathbf{x}) = \sqrt{v_{\eta'}(\mathbf{x})}$ ($\hat{s}_{\eta'}(\mathbf{x}) = \sqrt{\hat{v}_{\eta'}(\mathbf{x})}$).

The unit of the electron scattering intensity in the reconstruction at 3 days is set by the scaling described in Section 3.1.2 Item 3. The standard deviation has the same unit. The reconstructions at different time points, denoted by $\hat{\rho}(\mathbf{x})$, are scaled to the reconstruction at 3 days, denoted by $\hat{\rho}^{\text{capsid}}(\mathbf{x})$, by the following algorithm. First, compute the optimal gain g_* by $g_* = \arg \min_g \|g\hat{\rho}(\mathbf{x}) - \hat{\rho}^{\text{capsid}}(\mathbf{x})\|$ where $\|f\| = \int |f(\mathbf{x})|d\mathbf{x}$. Second, the scaled reconstruction is $g_*\hat{\rho}(\mathbf{x})$.

Figure 3.6 concerns variability versus time. Variability is described by averaged standard deviation and is computed as follows. Let $v_{\eta',\delta}(\mathbf{x})$ be the variance for the δ th repetition of the calculation. Define

$$\bar{s}_{\eta',\delta} = \sqrt{\int_{\mathbf{x} \in \Upsilon} v_{\eta',\delta}(\mathbf{x})d\mathbf{x} / \int_{\mathbf{x} \in \Upsilon} 1d\mathbf{x}}. \quad (3.33)$$

Then the plotted value is

$$\bar{\bar{s}}_{\eta'} = \frac{1}{\Delta} \sum_{\delta=1}^{\Delta} \bar{s}_{\eta',\delta} \quad (3.34)$$

and the sample standard deviation marks are at $\pm\mu_{\eta'}$ where

$$\mu_{\eta'} = \sqrt{\frac{1}{\Delta} \sum_{\delta=1}^{\Delta} [\bar{\bar{s}}_{\eta'} - \bar{s}_{\eta',\delta}]^2}. \quad (3.35)$$

For the capsid calculation, the volume Υ is the annulus with inner radius 120Angstrom and outer radius 216Angstrom. For the four subunit calculations, the volume Υ is described implicitly by the following algorithm: (1) Compute a cube of the space-varying variance map with sampling interval 2.768Angstrom. (2) Rotate the cube from the coordinate system of Ref. [85] to the coordinate system of VIPERdb [77]. (3) In the refined crystal structure for N ω V (1OHF) [24,45], locate all amino acids for which the α carbon is within 10Angstrom of the α carbon of the asparagine at the cleavage site (Asn570). (4) Locate a $3 \times 3 \times 3$ cube of voxels around each voxel containing a α carbon in Step (3). This collection of voxels is the volume denoted by Υ .

Spherical averages of the variance map are used in Figure 3.6(D–E). Each spherical average is computed using a formula analogous to Ref. [83, Eqs. 22–25], specifically,

$$\bar{v}_{\eta'}(\mathbf{x}) = \frac{1}{4\pi} \int v_{\eta'}(\mathbf{x}) d\Omega \quad (3.36)$$

$$= \frac{1}{4\pi} \sum_{l=0}^L \sum_{p=1}^P \left[\sum_{n=0}^{N_l-1} v_{l,n,p}^{(\eta')} \left(\sum_{m=-l}^{+l} |b_{l,n,m}|^2 \right) \right] h_{l,p}^2(x) \quad (3.37)$$

where $(V_{\eta'})_{j,j'} = v_{l(j),n(j),p(j)}^{(\eta')} \delta_{j,j'}$, $h_{l,p}(\cdot)$ is the radial basis function [83], $\int d\Omega$ is integration over the sphere, and

$$I_{l,n}(\theta, \phi) = \sum_{m=-l}^{+l} b_{l,n,m} Y_{l,m}(\theta, \phi) \quad (3.38)$$

where $I_{l,n}(\cdot, \cdot)$ is the (l, n) th icosahedral harmonic [85] and $Y_{l,m}(\cdot, \cdot)$ is the (l, m) th spherical harmonic. Applying Eq. 3.37 to the results of multiple calculations on different data sets indexed by $\delta \in \{1, \dots, \Delta\}$ gives $\bar{v}_{\eta', \delta}$. Then, $\bar{s}_{\eta', \delta} = \sqrt{\bar{v}_{\eta', \delta}}$. Finally, the sample mean and sample standard deviations of the spherical averages are computed by Eqs. 3.34 and 3.35.

3.1.3 Results

Figure 3.3 shows the four time-resolved reconstructions as surface and cross section plots. These plots are colored by the square root of the variance map (i.e., the standard deviation map). The overall impression from the capsid surfaces shown in Figure 3.3(A) is that the variability decreases in amplitude as time passes and the particle matures. The gradual stabilization of the capsid can be easily appreciated by comparing the variance at 3 minutes, 30 minutes and 4 hours time points. However, if individual scales are used to plot the variance map, it became apparent that the stabilization process is still incomplete 4 hours after the initiation of maturation (Figure 3.3(B)). Because the variance is computed for each voxel of the reconstruction, we can analyze the stabilization process for the entire structure, as demonstrated by the cross-section view in Figure 3.3(B). It can be seen that even 3 days after maturation the internal density continues to have high variance, which is expected if the RNA core of the particle is not highly ordered and does not obey icosahedral symmetry. However, the protein shell of the completely cleaved particle, i.e., the infectious particle, still retains a region of relative high variance in the center of the five fold axes.

Figure 3.5 provides information about resolution as Fourier Shell Correlation (FSC) plots for the reconstructions. In the first column of Figure 3.5, the FSC plots show that the achieved resolutions for the reconstructions are 22, 21, 21, and 20 Angstrom for 3 minutes, 30 minutes, 4 hours, and 3 days, respectively. In the second column of Figure 3.5, the FSC plots show that the reconstructions at times 3 minutes, 30 minutes, and 4 hours agree with the reconstruction at time 3 days within resolutions of 27, 24, and 33 Angstrom, respectively. In all cases, resolution is defined to be the inverse spatial frequency when the curve first intersects 0.5. Resolution is also described in Figures 3.10 and 3.11. Figure 3.10 shows the 3 minute reconstruction at reduced resolution, specifically, using only 180 coefficients (the so-called Step 5 of Ref. [83]) and Figure 3.11 shows cross sections of the four reconstructions colored by the mean map or colored by the square root of the variance map (i.e., the standard deviation map).

Next, we used the reconstructions generated in this work to analyze the variance around the cleavage site in each of the four quasi-equivalent subunits. Figure 3.6(A) shows the position of subunits A, B, C, and D in the $T = 4$ surface lattice with the location of the autocatalytic site indicated by a red cross. The voxels covering the region occupied by amino acid residues within 10 Angstrom of the active site were used to quantify the variance around the autocatalytic site of each subunit (Figure 3.6(B)). This is the same region analysed by Matsui *et al.* (2010) with difference maps. Figure 3.6(C) shows that the variance in volumes encompassing the B and C active sites is clearly higher than the variance in A and D active sites at early time points and they all converge to the same variance at later time points when all the subunits have cleaved. We assume that positions of higher variance are still changing and that these cleavage sites have not yet formed. This is consistent with the position-specific active site formation ob-

served by Matsui *et al.* (2010), validating the maximum-likelihood derived variance maps as a quantitative tool to address protein dynamics. An unexpected feature observed in this new analysis is that not only the active sites but also the average of the annulus containing the majority of the capsid protein densities reduce in variance by at least a factor of 3 as all of the subunits undergo cleavage. This important result could not be determined from the analysis of difference maps, but emerges naturally from the time resolved data sets when analyzed with the maximum likelihood algorithm that explicitly takes into account the continuous variability from one instance to another instance of the particle. Figure 3.6(D-E) shows the radial dependence of the average of the variance map for each of the four time points, which decreases by a factor of 4 from time 3 minutes to time 3 days independent of the radial position from the center of the particle. All standard deviations plotted in Figure 3.6 are computed by Eq. 3.35 with $\Delta = 4$ repeats of the reconstruction based on nonoverlapping subsets of the image stack at a particular time point.

Figure 3.9 shows ribbon diagrams of each of the subunits at each of the time points colored by the standard deviation map. The standard deviation tends to be largest in the helical region of the capsid protein near the autocatalytic site (Asn 570). These diagrams emulate diagrams used to display the Debye-Waller temperature factor in crystallography, where similar plots are made with the temperature factor displayed using color for each alpha carbon position of the peptide. The coordinates shown in the ribbon diagrams are from the refined x-ray crystallographic structure of N ω V (1OHF) [24, 45]. The standard deviation of the pixel position closest to a given C α coordinate was used to color code the ribbons.

3.1.4 Discussion

We showed that the maximum likelihood derived variance maps calculated for $N\omega V$ in different stages of maturation successfully captured the same subunit-specific dynamics features previously observed with difference maps by Matsui, Lander, Khayat, and Johnson [41]. However, while the difference map analysis was technically limited to a small portion of the structure, this new approach allowed us to observe an overall reduction in structural variability as the particle matures. This data correlates with the increase in particle stabilization as a function of cleavage, as previously demonstrated biochemically [66]. Moreover, this new analysis afforded the identification of highly dynamic regions in the fully mature capsid that were not obvious in the crystal structure. At the 5-fold symmetry axes, the high variance region (Figure 3.3) encompasses a central channel formed by the C-terminal gamma peptide of Subunit A and the N-terminal helix of Subunit B (Figure 3.9). Recently, it was demonstrated that $N\omega V$ membrane disruption activity is promoted by gamma peptides specifically derived from Subunit A cleavage [17]. In the crystal structure the five-fold central channel is protected from the solvent, however, the increased mobility observed in this analysis agrees with the high dynamics required to expose gamma peptide to the external environment, where it would be accessible to protease activity, as already demonstrated in Ref. [9], and could interact with cellular membranes. Therefore, the maximum likelihood derived variance maps can possibly provide information about putative binding sites and regulatory regions in cryo-EM structures. Another important advantage of the approach proposed here is that no difference maps are involved so the method would still be applicable if the overall structure underwent large changes.

The method described in this thesis is based on simultaneously computing a nominal reconstruction and a map of the space-varying heterogeneity of a biological particle from single-particle cryo EM data. The method depends on describing the heterogeneity probabilistically and estimating the statistics of the heterogeneity from the image data. This is a generalization of previous work [16,34,36,51,57,83] to the case where the particle is described probabilistically rather than deterministically. The method can be extended from maximum likelihood estimation to maximum *a posteriori* estimation (which is of interest in the biology community [56]) as is described for the homogeneous case in Ref. [16, Section VII]. In this chapter, resolution is measured by the standard FSC method of comparing two reconstructions computed from non-overlapping sets of images. This can be done rapidly using previously published formulas [83, Eqs. 22–25]. A more statistical approach that is natural for maximum likelihood estimators has been described [51, Section 4]. In the approach proposed here for heterogeneous particles, both of these methods measure resolution in terms of the nominal structure not the variance map.

The resolutions of the maps presented here are moderate compared to the sub nanometer reconstructions in Matsui, Lander, Khayat, and Johnson [41] due to the computationally intensive nature of the algorithm and current limited computing capability. In the $N\omega V$ example of Section 3.1.3, the new method produced variance maps that agreed closely with the difference maps computed at the higher resolution emphasizing the power of the method and motivating the use of high performance computers that will allow calculations to be performed at the resolutions dictated by the data. Potentially, though not demonstrated in the example of this chapter, the simultaneous estimation will lead to a better reconstruction since the reconstruction algorithm is allowed the additional

degrees of freedom of assigning high variance to a part of the structure rather than allowing the somewhat disordered state of a segment of the structure to contaminate better ordered parts of the structure. The method presented in this thesis should have broad application to existing EM data sets that can be reanalyzed with explicit spatial variance maps that may well provide added value for relating these structures to the function of the macromolecules.

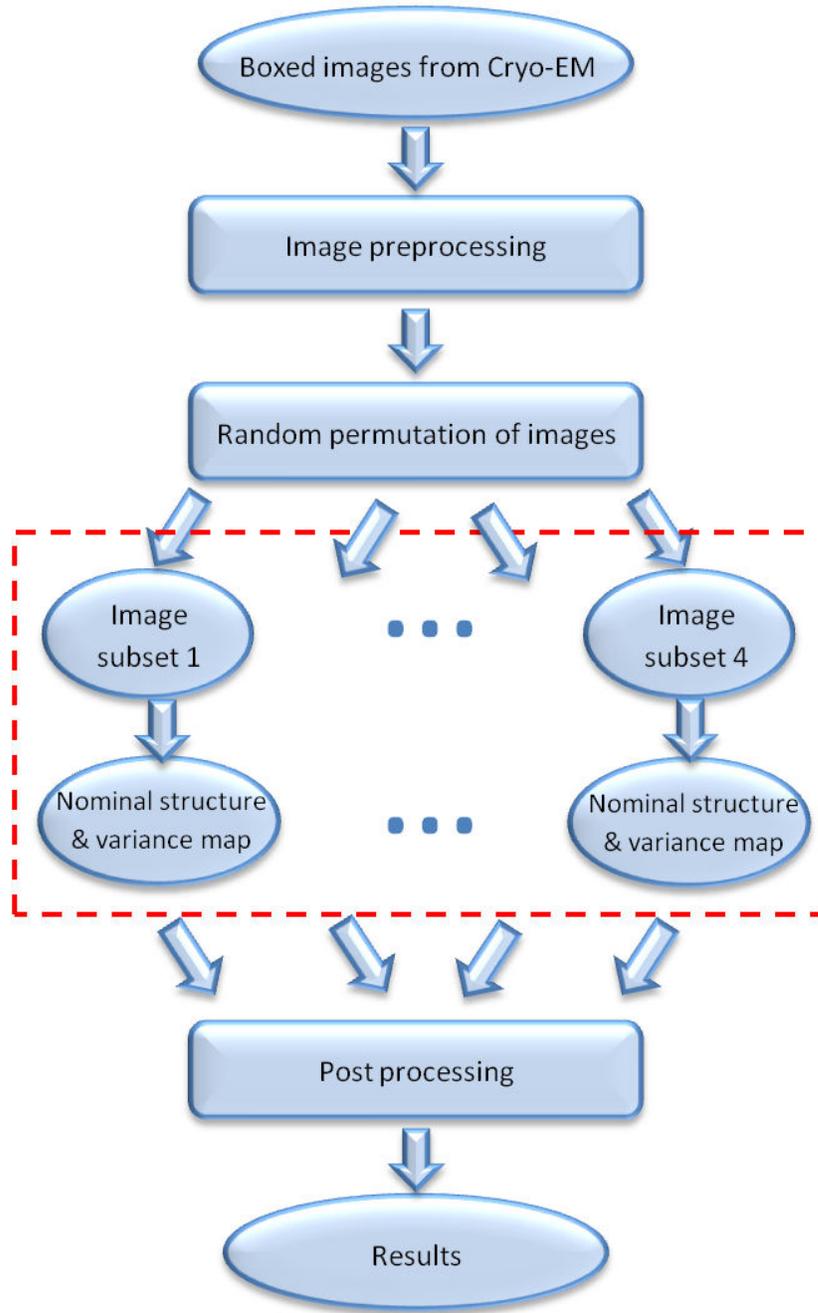


Figure 3.1: Algorithm flowcharts. The four parallel computations in Panel (a) are used to determine the performance of the algorithm, e.g., the error bars in Figure 3.6. The calculations contained in the red dotted-line box are expanded in Figure 3.2.

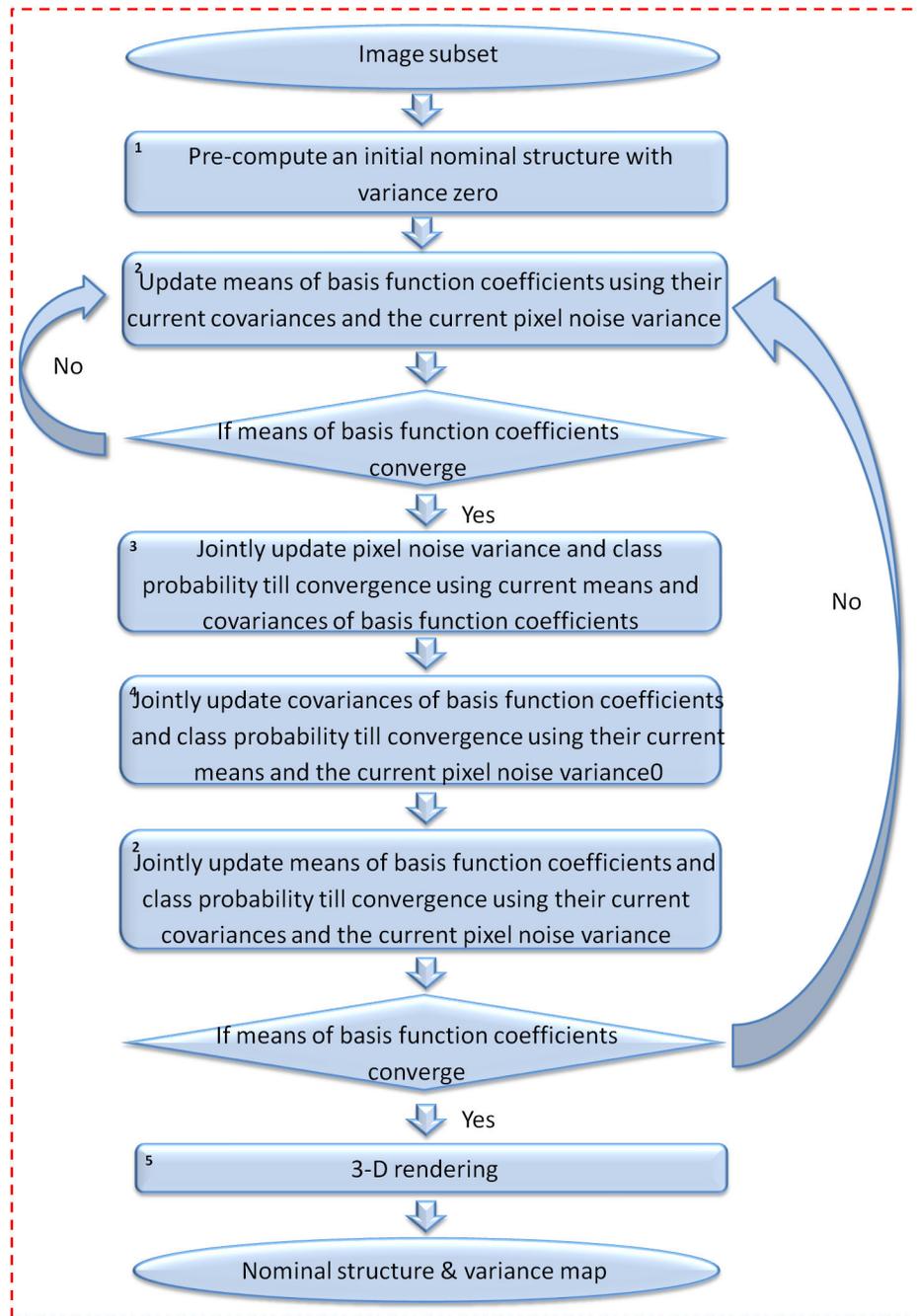


Figure 3.2: Algorithm flowcharts expanded. This is an expanded view of the red dotted-line box of Figure 3.1 which describes the maximum likelihood estimator.

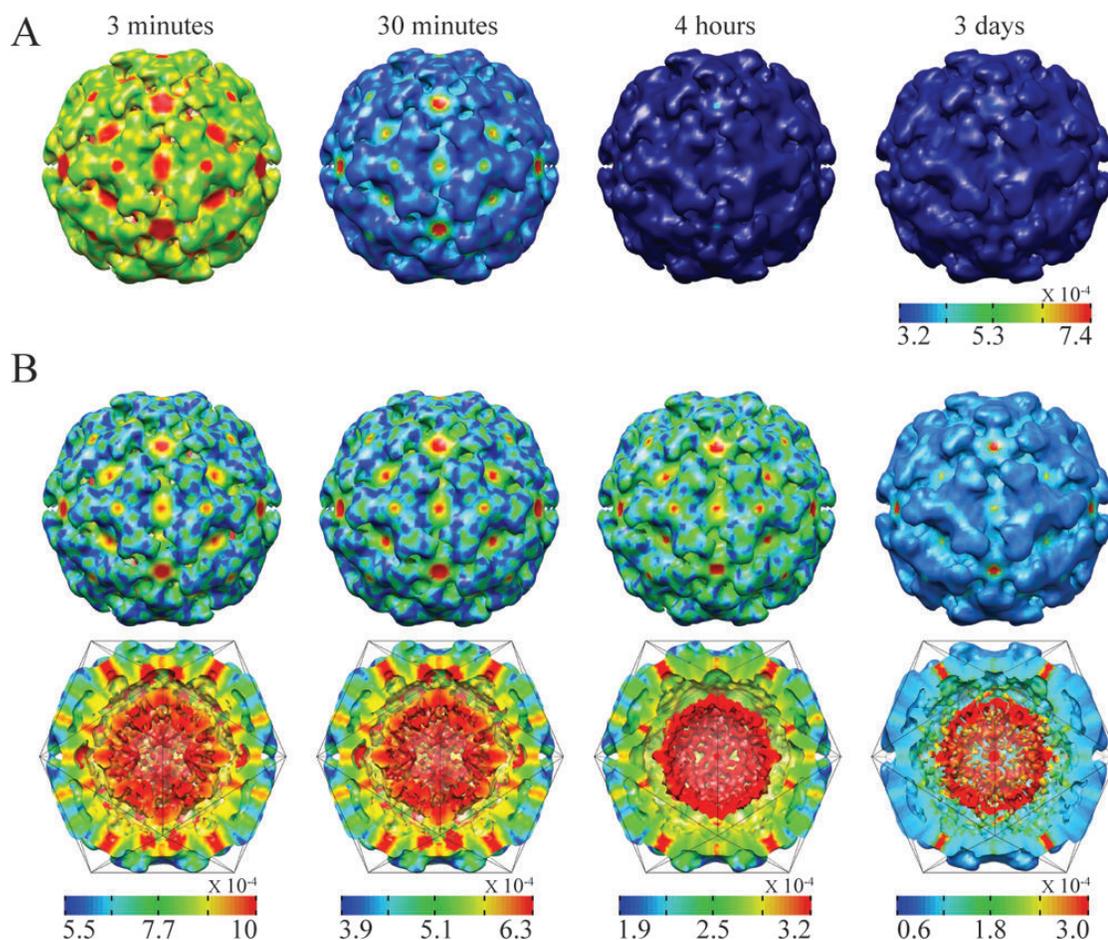


Figure 3.3: The four time-resolved reconstructions. Panel A: Surface of each of the four reconstructions colored by the square root of the variance map (i.e., the standard deviation map) and displayed using the VIPERdb [77] convention. The same color map is used in all images. Panel B: The surface and a cross section perpendicular to a 2-fold axis of each of the four reconstructions colored by the standard deviation map. The surface and cross section visualizations at a particular time point share the same color map. Different color maps are used at different time points. Visualization by UCSF Chimera [49].

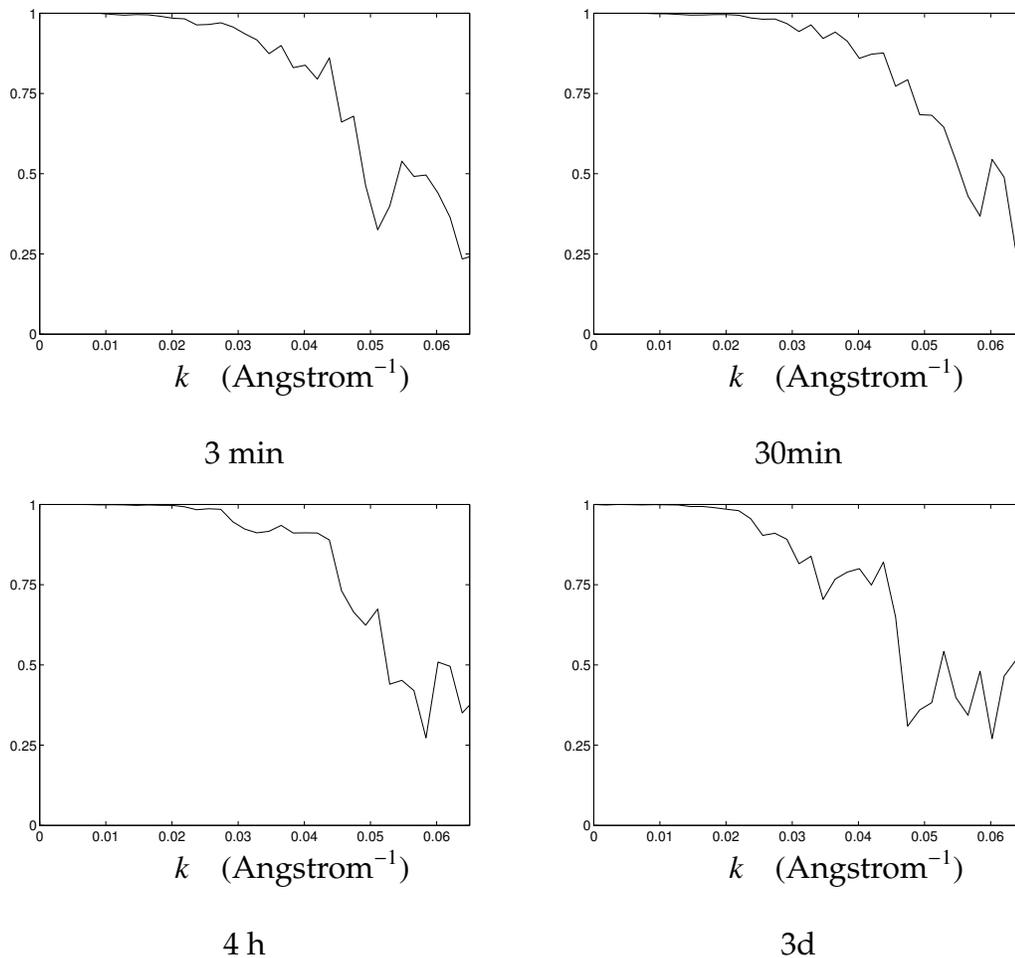


Figure 3.4: Part I of the resolution of the four time-resolved reconstructions as a function of k , which is the magnitude of the reciprocal-space frequency vector measured in \AA^{-1} . Fourier Shell Correlation (FSC) curves for comparing reconstructions from non-overlapping subsets containing 1200 images from the same data set. Based on these curves, the resolution of the four structures are approximately 21 \AA . All FSC curves were computed using command `proc3d` in EMAN.

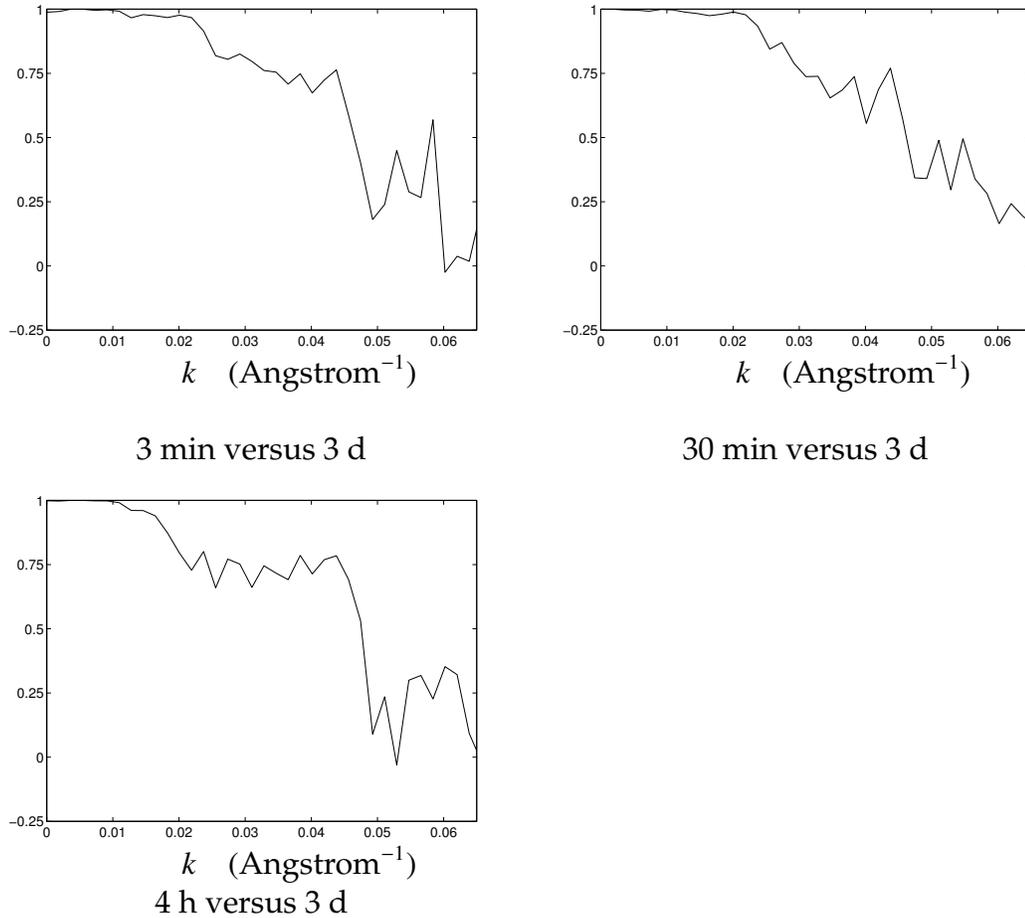


Figure 3.5: Part II of resolution of the four time-resolved reconstructions as a function of k , which is the magnitude of the reciprocal-space frequency vector measured in Angstrom^{-1} . Fourier Shell Correlation (FSC) curves between the 3 days reconstruction and each of the 3 minutes, 30 minutes, and 4 hours reconstructions for the nominal structures. Based on these difference curves, all the early structures agree with the capsid structure to approximately 24–33 Angstrom. All FSC curves were computed using command `proc3d` in EMAN.

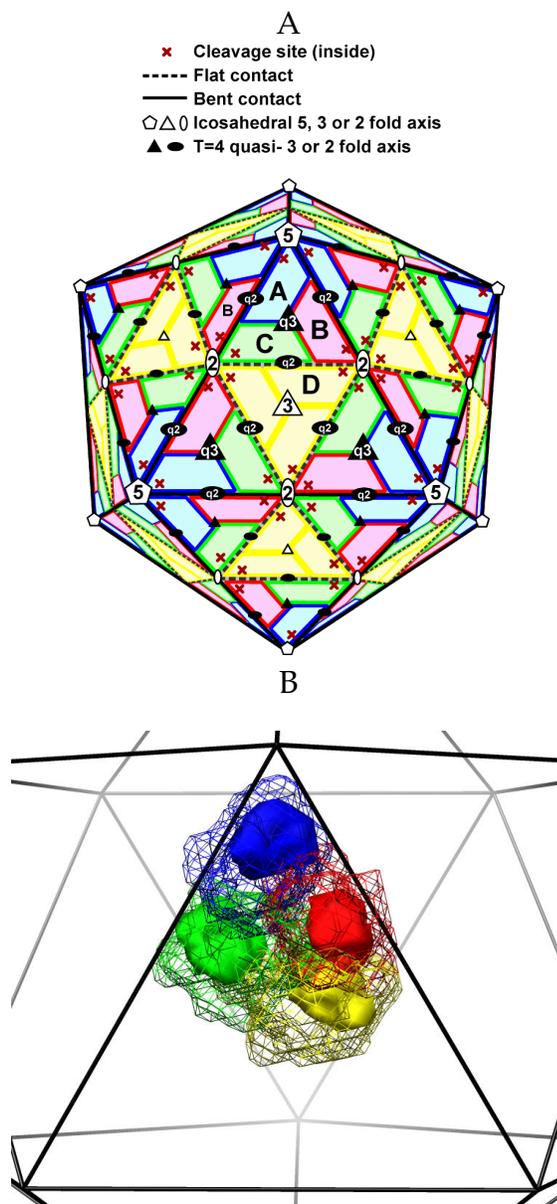


Figure 3.6: Part I of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. Panels A and B: Variance analysis around the cleavage sites of Subunits A, B, C and D that form the asymmetric unit of the NwV protein capsid. Both panels show the $T = 4$ surface lattice with the subunits' locations. The total volume occupied by each subunit is rendered as a mesh in Panel B. The variance was calculated over a smaller region, enclosing the cleavage site, which is shown as a solid volume within the subunit density. As is described in Section 3.1.2, the smaller region is essentially the region occupied by $C\alpha$ atoms within 10Angstrom of the active site. This is the same region analyzed by Matsui *et al.* [41] using difference maps.

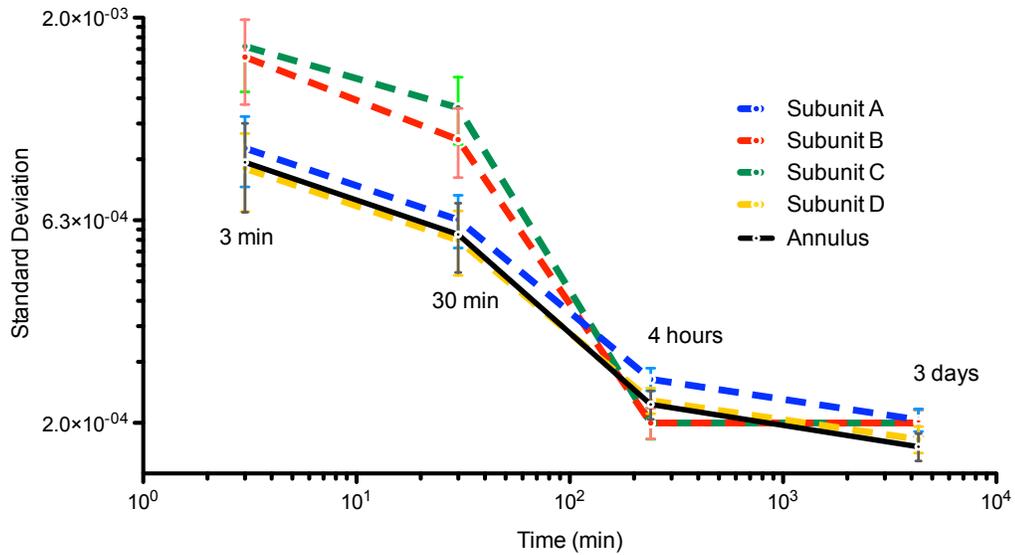


Figure 3.7: Part II of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. The standard deviation for the regions displayed in Panel B of Figure 3.6 are plotted log-log as a function of time for each subunit. The plot demonstrates an overall reduction of variance as a function of time after maturation is initiated, with distinct kinetics between the variances of the B and C sites (high) and the A and D sites (low). Computational methods are described by Eqs. 3.33, 3.34, and 3.35. The capsid shell is defined to be the annulus with radius from 120Angstrom to 216Angstrom.

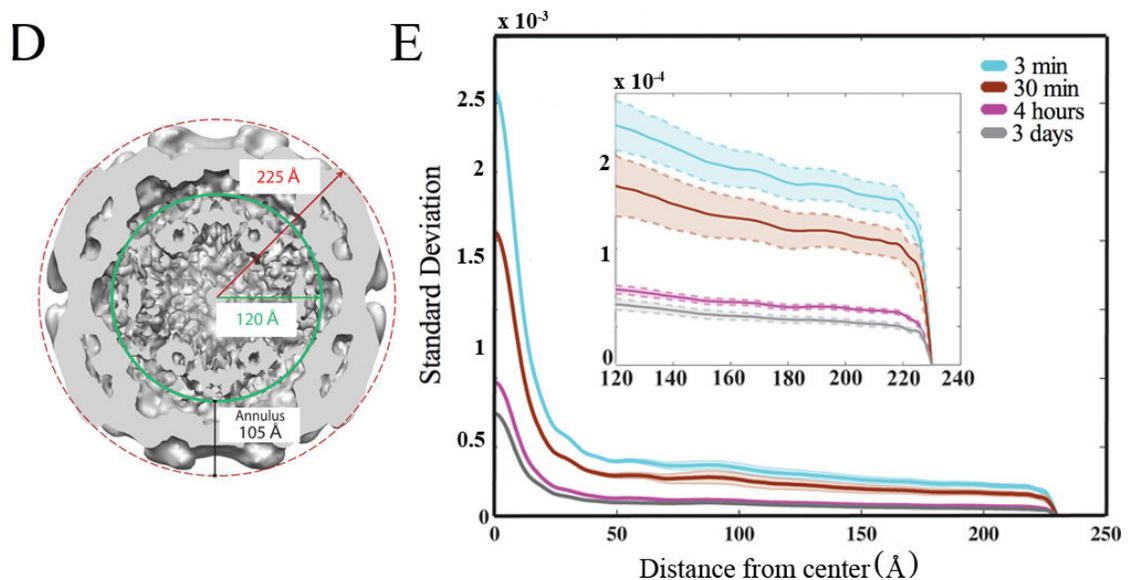


Figure 3.8: Part III of Region-specific variability analysis of the NwV protein capsid in different stages of maturation. Panels D–E: Time variation of spherical averages. A cross section perpendicular to the 2-fold axis (Panel D) shows the location of the capsid shell relative to the center of the particle. The square root of the spherically-averaged variance map versus distance from the center of the particle was computed by Eqs. 3.37, 3.34, and 3.35 and is plotted in Panel E. The shaded region covers plus/minus one standard deviation. The inset plot shows a zoomed version of the plot including only the capsid shell region.

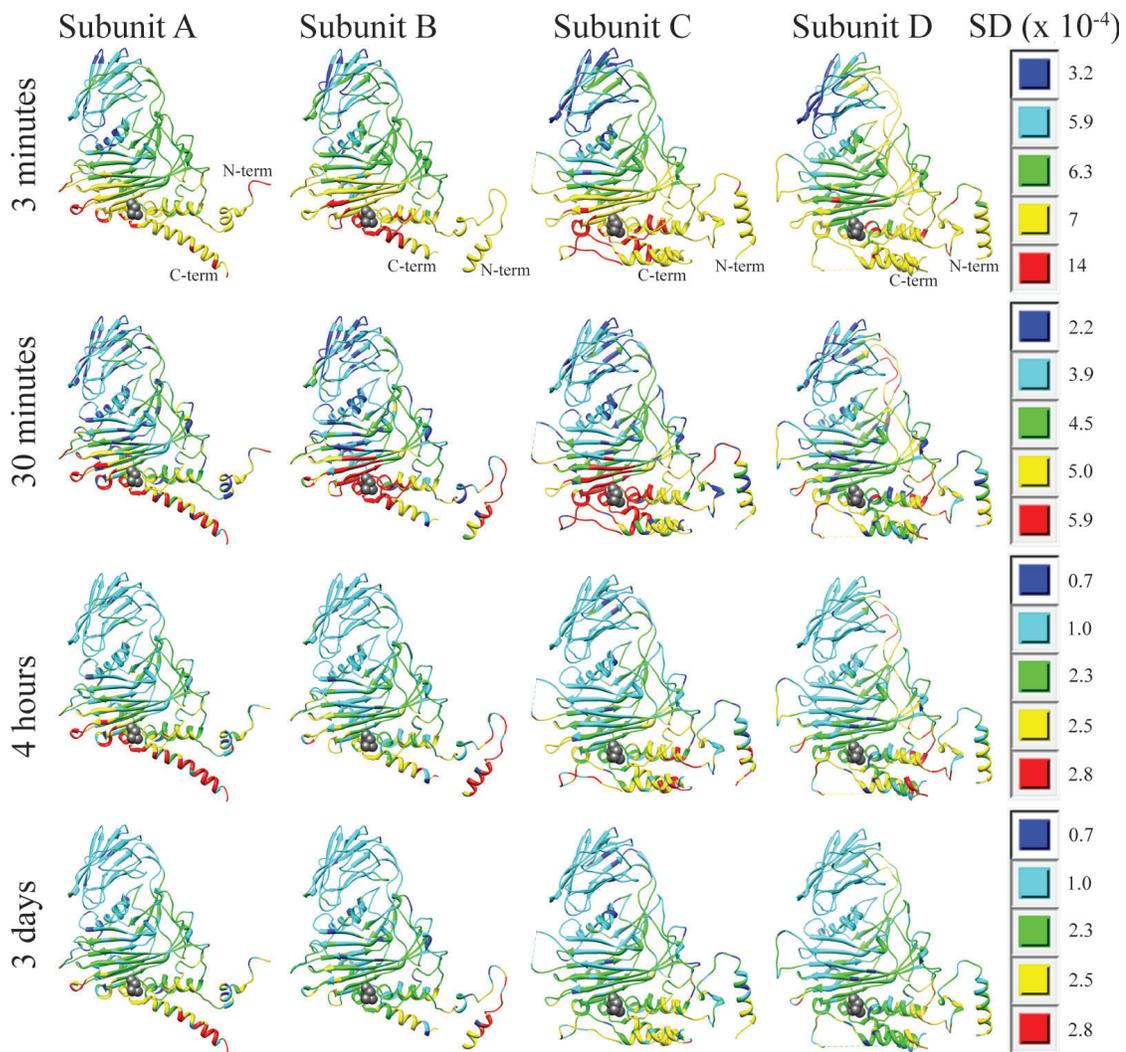


Figure 3.9: Ribbon diagrams of the four subunits at the four times colored by the square root of the variance map (i.e., the standard deviation map) with the asparagine at the self-catalytic site (Asn 570) shown as a ball-and-stick model. Each time point has its own color map analogous to the second row of Figure 3.3. For instance, red at the 3 minute time point is $14 \times 10^{-4} / 5.9 \times 10^{-4} = 2.4$ times higher than red at the 30 minute time point.

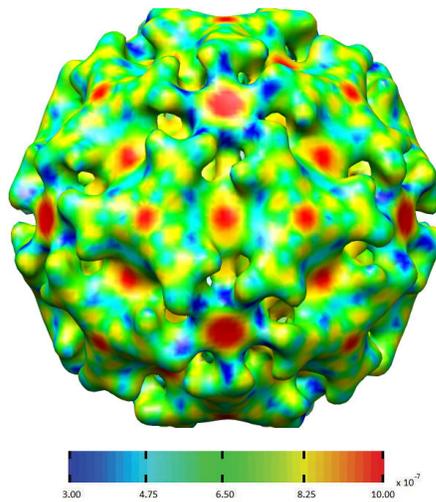


Figure 3.10: Surface of the 3 minute reconstruction colored by the square root of the variance map for a lower resolution reconstruction using 180 coefficients instead of 720 coefficients as was used in Figure 3.3 (the so-called Step 5 versus Step 7 of Ref. [83]). Visualization by UCSF Chimera [49].

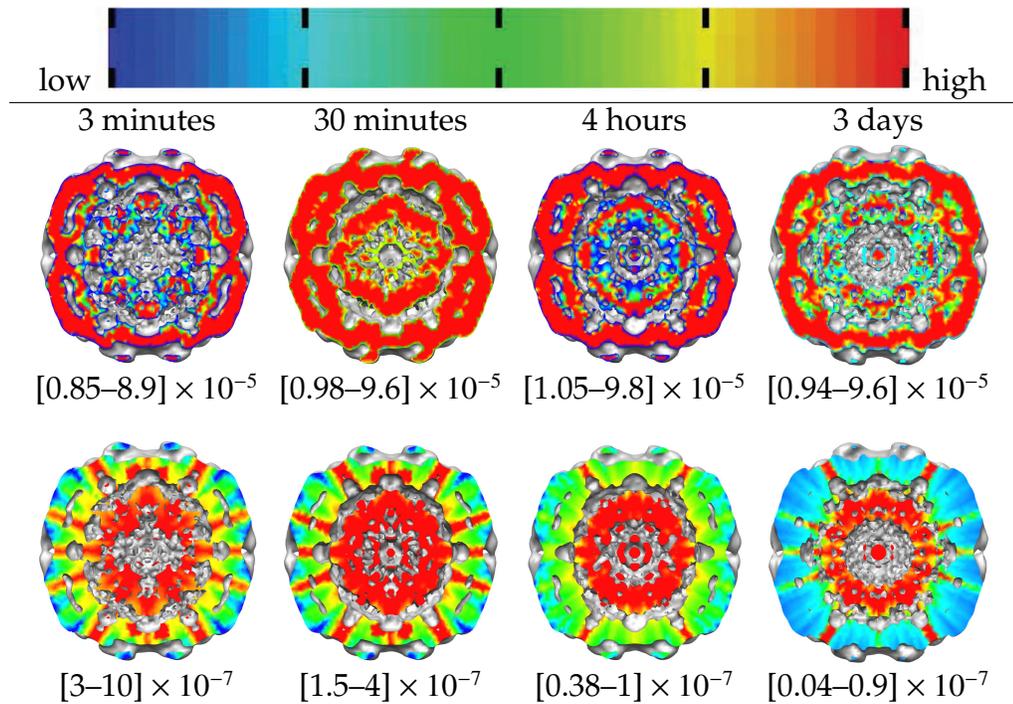


Figure 3.11: Cross sections perpendicular to a 2-fold symmetry axis colored by the mean map (first row) or by the square root of the variance map (second row). The mean map is roughly binary while the variance map has substantial spatial variation. Visualization by UCSF Chimera [49].

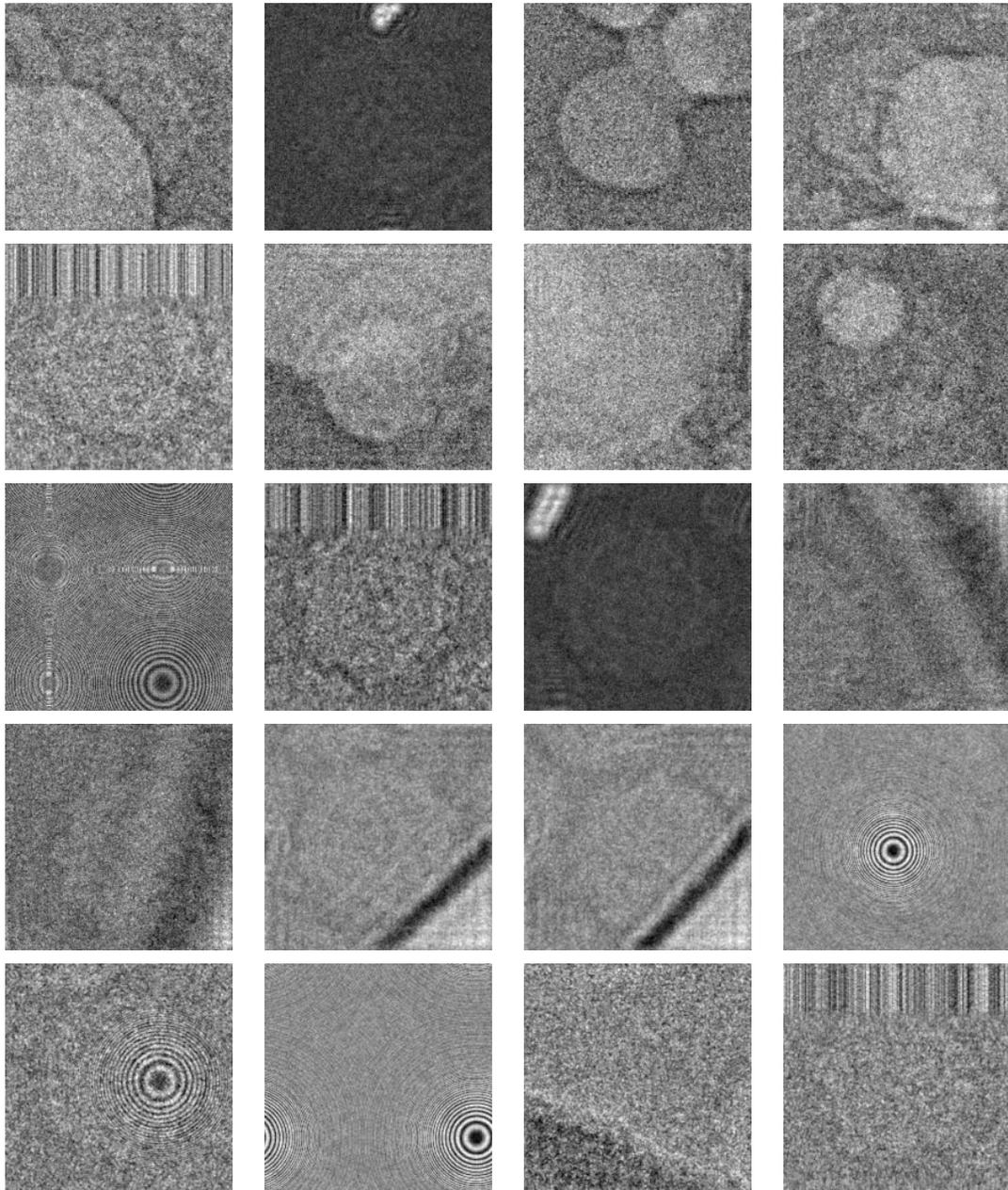


Figure 3.12: Examples of images from the 3 minute image stack that were rejected by the algorithm of Section 3.1.2. These images were excluded from the reconstruction calculations.

3.2 Other applications

Heterogeneous reconstruction calculations using the ideas of this thesis have been done on two additional viruses. Preliminary results are summarized in the following two sections.

3.2.1 Heterogeneity of NT procapsid, NT capsid and WT mature capsid of the $N\omega V$ particles

A mutant version of the $N\omega V$ virus (named NT) has been developed in which the cleavage reaction of the wild type particle has been blocked. In this work, the ideas of this thesis are applied to three previously processed cryo-EM data sets: NT procapsid, NT capsid and wild-type (WT) mature capsid which is constructed of cleaved peptides. The heterogeneous reconstruction results reveal the intrinsic heterogeneity of the three particles with the NT capsid having the highest variance, the mature cleaved WT capsid the lowest variance and the NT procapsid an intermediate level of variance. Mapping the variance to four subunits in the virus asymmetric unit provides a clear pattern of structural flexibility within the subunits in different particles. Compared with the WT capsid, the non-cleaving NT capsid reconstruction had disordered regions in the N-terminal helical bundle around the 5-fold and C-terminal switch helices. Consistent with the biochemical observations, the new variance analyses demonstrate a significant increase in flexibility in the disordered regions in the NT capsid compared with the mature cleaved WT capsid [64, 65].

3.2.2 Heterogeneous *Hong Kong 97 Virus* (HK97)

As is described in Chapter 5, the computational complexity of the algorithms described in this thesis is a challenge. The simplest approach is to gain access to a larger computer that has Matlab [40]. Following this approach, the software has been ported to the Gordon XSEDE cluster at the San Diego Supercomputer Center.

Two sets of images of the bacteriophage Hong Kong 97 (HK97) [25, 27] are available and represent a suitable more challenging problem in comparison with the N ω V images. For N ω V, the particle radius is 216Angstrom, 1200 images are used, in each image 8000 reciprocal space pixels are used, and 720 basis functions are used. For HK97, the particle radius is 280Angstrom which is larger, 1200 images are used, in each image 8000 reciprocal space pixels are used, and 900 basis functions are used because of the larger particle radius.

The biological question is the effect of the protease on the heterogeneity of the capsid. In the maturation of this bacteriophage, the approximately 60 copies of the protease peptide cleave the 7×60 copies of the capsid peptide and then the protease and the smaller of the cleavage products diffuse out of the particle. In preliminary calculations, the heterogeneous reconstruction algorithms of this thesis indicate that the particle with protease and before cleavage has greater heterogeneity (larger variance) than the particle without protease and after cleavage. Further calculations will be necessary in order to achieve higher spatial resolution and in order to spatially locate the regions of highest variance relative to the geometry of the particle.

CHAPTER 4
RECIPROCAL SPACE REPRESENTATIONS OF HELICAL-LIKE OBJECTS
WITH INFINITE PERIOD

4.1 Introduction

The 3-D Fourier transform of an object with helical symmetry is well known [14, 44]. Recently, an alternative formulation focused on objects having helical symmetry and constructed from repetitions of a motif was presented [35]. Biological objects often are only approximately helical. As described in this chapter, the motif approach allows the representations of objects having the combined rotation and translation symmetry of a helical object but with an infinite period, which previously could only be approximated by using large u and v indices.

Helical symmetry requires that the rotation around the z axis between repetitions is $\phi_0 = 2\pi v/u$ and the translation along the z axis between repetitions is $z_0 = c/u$ where c is the period of the helix and u and v are relatively prime integers. These conditions imply that the rotation angle $\phi_0/(2\pi) = v/u$ is a rational number and that the period $c = z_0 u$ is proportional to u . In the generalization described in this chapter, ϕ_0 and z_0 are unrestricted, i.e., arbitrary real numbers. By using large values of u and v , a rational approximation of the real valued $\phi_0/(2\pi)$ can be determined leading to a large period $c = z_0 u$ but this approximation is exact only in the limit as u and v grow infinitely large (and therefore c also grows infinitely large) and large u and v may not be attractive from a computational point of view.

The results in this chapter are based on two equations of Ref. [35]. From [35,

Eq. 5], the electron scattering intensity (denoted by $\rho_H(\cdot)$) of the helix represented by an array of motifs (denoted by $\rho_M(\cdot)$) is

$$\rho_H(\mathbf{x}) = \sum_{j=-\infty}^{+\infty} \rho_M \left(R_H^{-1} \left[S_{j\phi_0}^{-1} (\mathbf{x} - jz_0 \mathbf{e}_z) - \mathbf{x}_H \right] \right). \quad (4.1)$$

If ρ_M and ρ_H have 3-D Fourier transforms P_M and P_H , then

$$P_H(\mathbf{k}) = \sum_{j=-\infty}^{+\infty} \exp(-i2\pi \mathbf{k}^T [S_{j\phi_0} \mathbf{x}_H + jz_0 \mathbf{e}_z]) P_M((S_{j\phi_0} R_H)^{-1} \mathbf{k}). \quad (4.2)$$

A key challenge solved in this chapter is to show the presence and pattern of the layer lines for this generalization of helical symmetry.

4.2 The case of spherically symmetric motifs

To simplify the presentation, we first consider motifs with spherical symmetry in which case $P_M(\mathbf{k})$ is a function of $|\mathbf{k}|$ only, which is denoted by $P_M(|\mathbf{k}|)$. Then $P_H(\mathbf{k})$ (Eq. 4.2) is simplified to

$$P_H(\mathbf{k}) = P_M(|\mathbf{k}|) \sum_{j=-\infty}^{+\infty} \exp(-i2\pi \mathbf{k}^T [S_{j\phi_0} \mathbf{x}_H + jz_0 \mathbf{e}_z]). \quad (4.3)$$

Let $\mathbf{k} = (k_x \ k_y \ k_z)^T$ in Cartesian coordinates, $\mathbf{k} = (k_0 \ \phi_k \ k_z)^T$ in cylindrical coordinates, and $\mathbf{k} = (|\mathbf{k}| \ \theta_k \ \phi_k)^T$ in spherical coordinates. It follows that

$$\mathbf{k}^T S_{j\phi_0} \mathbf{x}_H = r_H (k_x \cos j\phi_0 - k_y \sin j\phi_0) \quad (4.4)$$

$$= r_H k_0 (\cos \phi_k \cos j\phi_0 - \sin \phi_k \sin j\phi_0) \quad (4.5)$$

$$= r_H k_0 \cos(\phi_k + j\phi_0). \quad (4.6)$$

By substituting Eq. 4.6 into Eq. 4.3, it follows that

$$P_H(\mathbf{k}) = P_M(|\mathbf{k}|) \sum_{j=-\infty}^{+\infty} \exp(-i2\pi [r_H k_0 \cos(\phi_k + j\phi_0) + jz_0 \mathbf{k}^T \mathbf{e}_z]). \quad (4.7)$$

A key relationship is the generating function for Bessel functions (denoted by $J_\lambda(\cdot)$) which is [1, Eq. 9.1.41, p. 361]

$$\exp\left(\frac{1}{2}x\left(t - \frac{1}{t}\right)\right) = \sum_{\lambda=-\infty}^{+\infty} t^\lambda J_\lambda(x). \quad (4.8)$$

Using the substitutions $t = -i \exp(i(\phi_k + j\phi_0))$ and $x = 2\pi r_H k_0$ in Eq. 4.8, the generating function implies

$$\exp(-i2\pi r_H k_0 \cos(\phi_k + j\phi_0)) = \sum_{\lambda=-\infty}^{+\infty} \exp(i\lambda(\phi_k + j\phi_0 - \frac{\pi}{2})) J_\lambda(2\pi r_H k_0). \quad (4.9)$$

Using Eq. 4.9 in Eq. 4.7 leads to

$$\begin{aligned} P_H(\mathbf{k}) &= P_M(|\mathbf{k}|) \sum_{j=-\infty}^{+\infty} \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k + j\phi_0 - \frac{\pi}{2}) - i2\pi j z_0 \mathbf{k}^T \mathbf{e}_z) \quad (4.10) \\ &= P_M(|\mathbf{k}|) \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k - \frac{\pi}{2})) \sum_{j=-\infty}^{+\infty} \exp(-i2\pi(z_0 \mathbf{k}^T \mathbf{e}_z - \frac{\lambda\phi_0}{2\pi})j). \end{aligned} \quad (4.11)$$

A second key relationship, from the theory of Fourier series, is

$$\sum_{n=-\infty}^{+\infty} \delta(t - n) = \sum_{k=-\infty}^{+\infty} e^{-i2\pi kt}. \quad (4.12)$$

Using Eq. 4.12 in Eq. 4.11 gives the result

$$P_H(\mathbf{k}) = \frac{1}{|z_0|} P_M(|\mathbf{k}|) \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k - \frac{\pi}{2})) \sum_{\nu=-\infty}^{+\infty} \delta(\mathbf{k}^T \mathbf{e}_z - \frac{\lambda\phi_0}{2\pi z_0} - \frac{\nu}{z_0}). \quad (4.13)$$

When the object is a helix, i.e., u , ν , and c are finite, Eq. 4.13 is identical to the key formula of Ref. [14], where $2\pi z_0/\phi_0 = P$ which is the pitch of the helix.

4.3 The case of general motifs

The motif is represented as an orthonormal expansion in spherical coordinates [35],

$$\rho_M(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} h_{l,p}(|\mathbf{x}|) \Psi_{l,m}(\theta_{\mathbf{x}}, \phi_{\mathbf{x}}), \quad (4.14)$$

and its 3-D Fourier transform is [35]

$$P_M(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \Psi_{l,m}(\theta_{\mathbf{k}}, \phi_{\mathbf{k}}). \quad (4.15)$$

It can be shown (see Appendix) that Eq. 4.2 and Eq. 4.15 imply that

$$P_H(\mathbf{k}) = \frac{1}{|z_0|} \sum_{n=-\infty}^{+\infty} \exp(in(\phi_k - \frac{\pi}{2})) \sum_{n'=-\infty}^{+\infty} \delta(\mathbf{k}^T \mathbf{e}_z - \frac{n\phi_0}{2\pi z_0} - \frac{n'}{z_0}) \hat{P}_M^n(\mathbf{k}), \quad (4.16)$$

where

$$\hat{P}_M^n(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \hat{\Psi}_{l,m}^n(\theta_k, \frac{\pi}{2}), \quad (4.17)$$

which is the yz -plane of a linearly distorted motif in reciprocal space, and where

$$\hat{\Psi}_{l,m}^n(\theta_k, \frac{\pi}{2}) = \sum_{q=-l}^{+l} b_{l,m,q} Y_{l,q}(\theta_k, \frac{\pi}{2}) J_{n-q}(2\pi r_H k_0), \quad (4.18)$$

which is a great circle in the yz -plane of a linearly distorted angular basis function.

The general case described in Eq. 4.16 is similar to the special case where the motif has spherical symmetry which is described in Eq. 4.13. The key complication in Eq. 4.16 is that the 3-D Fourier transform of the motif is combined with the Bessel functions and the result is inside the inner summation.

In the helical case, i.e., $\phi_0 = 2\pi v/u$, $z_0 = c/u$ and u and v are relatively prime, it can be shown (see Appendix) that Eq. 4.16 is equivalent to Eq. 6 of Ref. [35].

4.4 Discussion

The key feature of both Eq. 4.13 and Eq. 4.16 is that these aperiodic objects continue to have layer lines in 3-D reciprocal space. However, unlike the helical case, the layer lines are doubly rather than singly indexed. From the point of

view of computational burden in reconstruction software or other applications, it is an open question whether (1) a single sum with large u, v indices versus (2) a double sum truncated by the decay in $\hat{P}_M^n(\mathbf{k})$ as $|n|$ grows is less computation. Helical biological objects often have imperfect helical symmetry and the generalization described in this chapter may be a useful alternative to approximation with large u and v values. An alternative direction is generalization with stochastic helical or motif parameters.

4.5 Appendix

Derivation of Eq. 4.16

For any motif, Eq. 4.2 gives the 3-D reciprocal space representation of the 3-D object. The rotation matrix R_H is a convenience, because it allows the use of a different coordinate system when describing the motif, but is not necessary, i.e., R_H being the identity matrix is sufficient and we assume $R_H = I_3$ in the remainder of the chapter.

Apply the same techniques used to transform Eq. 4.7 to Eq. 4.11 on Eq. 4.2, and use $R_H = I_3$ to get

$$P_H(\mathbf{k}) = \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k - \frac{\pi}{2})) \left[\sum_{j=-\infty}^{+\infty} \exp(-i2\pi(z_0 \mathbf{k}^T \mathbf{e}_z - \frac{\lambda\phi_0}{2\pi})j) P_M(S_{j\phi_0}^{-1} \mathbf{k}) \right]. \quad (4.19)$$

Use Eq. 4.14 and Eq. 4.15 to represent the motif and its 3-D Fourier transform.

Let $\mu = S_{j\phi_0}^{-1} \mathbf{k}$. Then $|\mu| = |\mathbf{k}|$, since $S_{j\phi_0}$ is a rotation matrix, and

$$\mu = (S_{j\phi_0})^{-1} \mathbf{k} = \begin{pmatrix} k_x \cos j\phi_0 - k_y \sin j\phi_0 \\ k_x \sin j\phi_0 + k_y \cos j\phi_0 \\ k_z \end{pmatrix} = \begin{pmatrix} k_0 \cos(\phi_k + j\phi_0) \\ k_0 \sin(\phi_k + j\phi_0) \\ k_z \end{pmatrix}. \quad (4.20)$$

Hence, the angles of μ in spherical coordinates are $\theta_\mu = k_z/|\mu| = \theta_k$ and $\phi_\mu = \phi_k + j\phi_0$. Then it follows that

$$P_M(S_{j\phi_0}^{-1} \mathbf{k}) = P_M(\mu) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \Psi_{l,m}(\theta_k, \phi_k + j\phi_0). \quad (4.21)$$

For each choice of l and m there is a set of coefficients $b_{l,m,q}$ such that

$$\Psi_{l,m}(\theta, \phi) = \sum_{q=-l}^{+l} b_{l,m,q} Y_{l,q}(\theta, \phi) \quad (4.22)$$

$$= \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta) \exp(iq\phi). \quad (4.23)$$

Hence,

$$P_M(S_{j\phi_0}^{-1} \mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq(\phi_k + j\phi_0)). \quad (4.24)$$

Use Eq. 4.24 in Eq. 4.19 to get

$$P_H(\mathbf{k}) = \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k - \frac{\pi}{2})) \left[\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \times \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq\phi_k) \sum_{j=-\infty}^{+\infty} \exp(-i2\pi(z_0 \mathbf{k}^T \mathbf{e}_z - \frac{(\lambda + q)\phi_0}{2\pi})j) \right]. \quad (4.25)$$

Apply Eq. 4.12 to further simplify the formula to the form

$$P_H(\mathbf{k}) = \sum_{\lambda=-\infty}^{+\infty} J_\lambda(2\pi r_H k_0) \exp(i\lambda(\phi_k - \frac{\pi}{2})) \left[\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \times \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq\phi_k) \sum_{n'=-\infty}^{+\infty} \delta(z_0 \mathbf{k}^T \mathbf{e}_z - \frac{(\lambda + q)\phi_0}{2\pi} - n') \right]. \quad (4.26)$$

Replace λ by $n = \lambda + q$ to get

$$\begin{aligned}
P_H(\mathbf{k}) &= \sum_{n=-\infty}^{+\infty} \sum_{n'=-\infty}^{+\infty} \delta(z_0 \mathbf{k}^T \mathbf{e}_z - \frac{n\phi_0}{2\pi} - n') \left[\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \times \right. \\
&\quad \times \left. \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq\phi_k) J_{n-q}(2\pi r_H k_0) \exp(i(n-q)(\phi_k - \frac{\pi}{2})) \right] \quad (4.27) \\
&= \frac{1}{|z_0|} \sum_{n=-\infty}^{+\infty} \exp(in(\phi_k - \frac{\pi}{2})) \sum_{n'=-\infty}^{+\infty} \delta(\mathbf{k}^T \mathbf{e}_z - \frac{n\phi_0}{2\pi z_0} - \frac{n'}{z_0}) \times \\
&\quad \times \left[\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq\frac{\pi}{2}) J_{n-q}(2\pi r_H k_0) \right]. \quad (4.28)
\end{aligned}$$

Eq. 4.28 is the 3-D reciprocal space representation of an object where ϕ_0 and z_0 are not described by u, v, c and where the motif is general, i.e., does not have spherical symmetry. Let

$$\hat{P}_M^n(\mathbf{k}) = \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \hat{\Psi}_{l,m}^n(\theta_k, \frac{\pi}{2}) \quad (4.29)$$

and

$$\hat{\Psi}_{l,m}^n(\theta_k, \frac{\pi}{2}) = \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq\frac{\pi}{2}) J_{n-q}(2\pi r_H k_0). \quad (4.30)$$

Using these definitions in Eq. 4.28 implies

$$P_H(\mathbf{k}) = \frac{1}{|z_0|} \sum_{n=-\infty}^{+\infty} \exp(in(\phi_k - \frac{\pi}{2})) \sum_{n'=-\infty}^{+\infty} \delta(\mathbf{k}^T \mathbf{e}_z - \frac{n\phi_0}{2\pi z_0} - \frac{n'}{z_0}) \hat{P}_M^n(\mathbf{k}). \quad (4.31)$$

Eq. 4.31 is Eq. 4.16 in the chapter.

Equivalence of Eq. 4.16 and [35, Eq. 6]

In this section, it is shown that when the object is a helix with finite u, v, c , then Eq. 4.28 is equivalent to Eq. 6 of [35]. When $\phi_0 = 2\pi v/u$ and $z_0 = c/u$, Eq. 4.28

becomes

$$\begin{aligned}
P_H(\mathbf{k}) &= \sum_{n=-\infty}^{+\infty} \exp(in(\phi_k - \frac{\pi}{2})) \sum_{n'=-\infty}^{+\infty} u \delta(ck_z - nv - n'u) \times \\
&\times \left[\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq \frac{\pi}{2}) J_{n-q}(2\pi r_H k_0) \right] \quad (4.32)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \left\{ \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq \frac{\pi}{2}) \times \right. \\
&\times \left. \left[\sum_{n=-\infty}^{+\infty} \sum_{n'=-\infty}^{+\infty} u J_{n-q}(2\pi r_H k_0) \exp(in(\phi_k - \frac{\pi}{2})) \delta(ck_z - [n'u + nv]) \right] \right\}. \quad (4.33)
\end{aligned}$$

To separate the delta functions from the rest of the equation, it is necessary to change the summation indices for the part of the equation inside the square brackets. By the theory of Linear Diophantine equations [62, Section 2.5, p. 44], for any relatively prime integers u and v , there exists a pair of integers β and β' such that $u\beta' = v\beta + 1$. This is equivalent to

$$u = \frac{v\beta + 1}{\beta'} \quad (4.34)$$

$$v = \frac{u\beta' - 1}{\beta}. \quad (4.35)$$

Define

$$t = n'u + nv. \quad (4.36)$$

Then,

$$t = n' \frac{v\beta + 1}{\beta'} + nv \quad (4.37)$$

which is equivalent to

$$t\beta' = (n'\beta + n\beta')v + n'. \quad (4.38)$$

Similarly,

$$t = n'u + n \frac{u\beta' - 1}{\beta} \quad (4.39)$$

which is equivalent to

$$t\beta = (n'\beta + n\beta')u - n. \quad (4.40)$$

Define

$$\xi = n'\beta + n\beta'. \quad (4.41)$$

Then

$$\begin{cases} t\beta' = \xi v + n' \\ t\beta = \xi u - n \end{cases} \quad (4.42)$$

which is equivalent to

$$\begin{cases} n = \xi u - t\beta \\ n' = -\xi v + t\beta' \end{cases}. \quad (4.43)$$

Substitute the new indices in Eq. 4.33 to get

$$\begin{aligned} P_H(\mathbf{k}) &= \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \left\{ \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\theta_k) \exp(iq\frac{\pi}{2}) \times \right. \\ &\quad \left. \times \left[\sum_{t=-\infty}^{+\infty} \sum_{\xi=-\infty}^{+\infty} u J_{\xi u - t\beta - q}(2\pi r_H k_0) \exp(i[\xi u - t\beta][\phi_k - \frac{\pi}{2}]) \delta(ck_z - t) \right] \right\} \quad (4.44) \end{aligned}$$

$$\begin{aligned} &= \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \left\{ \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\theta_k) \exp(iq\frac{\pi}{2}) \times \right. \\ &\quad \left. \times \left[\sum_{t=-\infty}^{+\infty} u \delta(ck_z - t) \sum_{\xi=-\infty}^{+\infty} J_{\xi u - t\beta - q}(2\pi r_H k_0) \exp(i[\xi u - t\beta][\phi_k - \frac{\pi}{2}]) \right] \right\}. \quad (4.45) \end{aligned}$$

Use the integral representation of the Bessel function and standard Fourier series results [35, Supplemental Material] to get

$$\begin{aligned} &\sum_{\xi=-\infty}^{+\infty} J_{\xi u - t\beta - q}(2\pi r_H k_0) \exp(i[\xi u - t\beta][\phi_k - \frac{\pi}{2}]) \\ &= \sum_{t'=0}^{u-1} \exp(-i2\pi[r_H k_0 \cos(\frac{2\pi t'}{u} + \phi_k) - \frac{(t\beta + q)t'}{u}]) \exp(iq(\phi_k - \frac{\pi}{2})) \quad (4.46) \end{aligned}$$

$$= \sum_{t'=0}^{u-1} \exp(-i2\pi[r_H k_0 \cos(\frac{2\pi t'}{u} + \phi_k) - \frac{t\beta t'}{u}]) \exp(iq\frac{2\pi t'}{u}) \exp(iq(\phi_k - \frac{\pi}{2})). \quad (4.47)$$

Use Eq. 4.47 in Eq. 4.45 and rearrange the order of summations to get

$$\begin{aligned}
P_H(\mathbf{k}) &= \sum_{t=-\infty}^{+\infty} u\delta(ck_z - t) \sum_{t'=0}^{u-1} \exp(-i2\pi[r_H k_0 \cos(\frac{2\pi t'}{u} + \phi_k) - \frac{t\beta t'}{u}]) \times \\
&\quad \times \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq[\frac{2\pi t'}{u} + \phi_k])
\end{aligned} \tag{4.48}$$

$$\begin{aligned}
&= \sum_{t=-\infty}^{+\infty} u\delta(ck_z - t) \sum_{t'=0}^{u-1} \exp(-i2\pi[r_H k_0 \cos(\frac{2\pi vt'}{u} + \phi_k) - \frac{t t'}{u}]) \times \\
&\quad \times \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq[\frac{2\pi vt'}{u} + \phi_k])
\end{aligned} \tag{4.49}$$

where the last equality holds because the same set of values are enumerated in a different order [35, Supplemental Materials].

Since

$$\frac{2\pi vt'}{u} + \phi_k = t' \phi_0 + \phi_k \tag{4.50}$$

it follows that

$$\begin{aligned}
P_H(\mathbf{k}) &= \sum_{t=-\infty}^{+\infty} u\delta(ck_z - t) \sum_{t'=0}^{u-1} \exp(-i2\pi[r_H k_0 \cos(t' \phi_0 + \phi_k) - \frac{t t'}{u}]) \times \\
&\quad \times \sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq[t' \phi_0 + \phi_k]).
\end{aligned} \tag{4.51}$$

Apply the steps from Eq. 4.2 to Eq. 4.19 in reverse order to get

$$\exp(-i2\pi[r_H k_0 \cos(t' \phi_0 + \phi_k) - \frac{t t'}{u}]) = \exp(-i2\pi \mathbf{k}^T [t' z_0 \mathbf{e}_z + S_{t' \phi_0} \mathbf{x}_H]). \tag{4.52}$$

From Eq. 4.15, it follows that

$$\sum_{l=0}^{\infty} \sum_{m=0}^{2l} \sum_{p=1}^{\infty} d_{l,m,p} (-i)^l H_{l,p}(|\mathbf{k}|) \sum_{q=-l}^{+l} b_{l,m,q} N_{l,q} P_{l,q}(\cos \theta_k) \exp(iq[t' \phi_0 + \phi_k]) = P_M((S_{t' \phi_0} \mathbf{R}_H)^{-1} \mathbf{k}) \tag{4.53}$$

so that

$$P_H(\mathbf{k}) = \sum_{t=-\infty}^{+\infty} \delta(z_0 k_z - t) \sum_{t'=0}^{u-1} \left[\exp(-i2\pi\mathbf{k}^T [t' z_0 \mathbf{e}_z + S_{t'\phi_0} x_H]) P_M((S_{t'\phi_0} R_H)^{-1} \mathbf{k}) \right] \quad (4.54)$$

which is identical to Eq. 6 of [35].

CHAPTER 5

COMPUTATIONAL PERFORMANCE OPTIMIZATION

The maximum likelihood estimators described in this thesis require substantial computation, primarily because of the numerical evaluation of numerical integrations over the possible projection directions of each image. In this chapter we describe several types of approaches for reducing computation which include reorganizations of the algorithm, use of *a priori* information about the reconstruction, and software engineering.

5.1 Algorithm improvements

5.1.1 Implications of the matrix inversion lemma

The Matrix Inversion Lemma [2, Eq. 3.1, p. 139] can be applied to the definition of Σ (Eq. 2.7) to get

$$\Sigma_i^{-1}(\theta_i, \eta', V_{\eta'}, Q_i) = Q_i^{-1} - Q_i^{-1} L_i(\theta_i, \eta') \left(L_i^T(\theta_i, \eta') Q_i^{-1} L_i(\theta_i, \eta') + V_{\eta'}^{-1} \right)^{-1} L_i^T(\theta_i, \eta') Q_i^{-1}. \quad (5.1)$$

Define

$$D_i(\theta_i, \eta') \doteq L_i^T(\theta_i, \eta') Q_i^{-1} L_i(\theta_i, \eta') \quad (5.2)$$

$$\Delta_{\eta'}(y, \bar{\alpha}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \doteq \sum_{i=1}^{N_y} \int_{\theta_i} D_i(\theta_i, \eta') p(\theta_i, \eta' | y_i, \bar{\alpha}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \quad (5.3)$$

$$b_i(\theta_i, \eta', y_i) \doteq L_i^T(\theta_i, \eta') Q_i^{-1} y_i \quad (5.4)$$

$$\beta_{\eta'}(y, \bar{\alpha}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \doteq \sum_{i=1}^{N_y} \int_{\theta_i} b_i(\theta_i, \eta', y_i) p(\theta_i, \eta' | y_i, \bar{\alpha}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \quad (5.5)$$

$$T_i(\theta_i, \eta') \doteq \left(D_i(\theta_i, \eta') + V_{\eta'}^{-1} \right)^{-1}. \quad (5.6)$$

Then Eq. 2.33 can be rewritten in the form

$$F^{\eta'} \bar{c}^{\eta'} = g^{\eta'} \quad (5.7)$$

where

$$\begin{aligned} F^{\eta'} &= \Delta_{\eta'}(y, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) \\ &\quad - \sum_{i=1}^{N_y} \int_{\theta_i} D_i(\theta_i, \eta') (D_i(\theta_i, \eta') + V_{\eta'}^{-1})^{-1} D_i(\theta_i, \eta') p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \end{aligned} \quad (5.8)$$

$$\begin{aligned} g^{\eta'} &= \beta_{\eta'}(y, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) - \\ &\quad \sum_{i=1}^{N_y} \int_{\theta_i} D_i(\theta_i, \eta') (D_i(\theta_i, \eta') + V_{\eta'}^{-1})^{-1} b_i(\theta_i, \eta', y_i) p(\theta_i, \eta' | y_i, \bar{c}^{\eta'}, {}_0V_{\eta'}, \mathcal{Q}) d\theta_i \end{aligned} \quad (5.9)$$

and M , which is the key quantity for computing $\partial Q_1 / \partial V_{\eta'}$, can be rewritten (in abbreviated notation) in the form

$$M = [(I - DT)(b - Dc)][(I - DT)(b - Dc)]^T - (I - DT)D. \quad (5.10)$$

As is described in Ref. [34], under appropriate conditions (which include a pre-transformation of the image data), the $N_c(\eta) \times N_c(\eta)$ matrix D depends on only two of the three Euler angles that describe the projection orientation and are included in θ . If, furthermore, the structure of the equations is such that D can be factored out of the integral over the Euler angle on which it does not depend, then a fast algorithm of the type described in Ref. [34] can be derived. These conditions are satisfied for Eq. 5.7 and so a fast algorithm is available. A weaker set of conditions is satisfied by Eq. 5.10, since $(I - DT)$ [respectively, $(I - DT)^T$] can be factored out to the left [respectively, right] but $(b - Dc)(b - Dc)^T$ remains to be integrated with respect to all three Euler angles, so an algorithm of the same type is available but it provides less performance gain.

5.1.2 Implications of Sylvester’s Determinant Theorem

Sylvester’s determinant theorem is the analog for computing determinants of the Matrix Inversion Lemma. Let $A \in \mathbb{R}^{n \times n}$ be full rank and let $U, V \in \mathbb{R}^{n \times m}$. Then, $\det(A + UV^T) = \det(A) \det(I + V^T A^{-1} U)$. Determinants arise in numerical optimization of Q_1 with respect to V (Eq. 2.30 or Eq. 2.35). The two key dimensions are the number of coefficients $N_c(\eta)$ used to describe the electron scattering intensity (Eq. 2.1) and the number of pixels N_y in an image. The dimension of Σ is N_y and, in the calculations of Section 2.7, $N_c(1) = 720$ while $N_y = 91^2$ so using Sylvester’s theorem in the computation of $\det \Sigma$ reduces computation substantially.

5.1.3 Data-driven numerical integration rules

In many applications of the ideas of Chapters 2 and 3, a homogeneous reconstruction will be available before the heterogeneous reconstruction is computed. In this section, a method to exploit this information as *a priori* information is described. The basic idea is to use the homogeneous reconstruction to approximately estimate the projection orientation of each image and then only perform the numerical integration in the region of the approximate orientation. By reducing the volume of the set over which the integral is performed, the computational burden will be reduced. The final part of this section (starting at Item 4d) reflects details of the current software used to perform the heterogeneous reconstruction calculations of Chapters 2 and 3.

1. A homogeneous reconstruction and a set of images (an “image stack”)

is provided by the biologist user along with some sense of the ability to orient the images in the stack using the homogeneous reconstruction, e.g., “ $\pm 5^\circ$ ”. The scale of initial calculations are $N_v = 1200$ images in the stack and $N_y = 8000$ reciprocal-space pixels in each image are actually used.

2. A projection orientation can be thought of as a direction in space (described by the angles of spherical coordinates, for instance) and a rotation around the direction. The location of the abscissas of an integration rule can be selected in the following manner:

- (a) The directions in space, or equivalently positions on the surface of a unit sphere, can be selected with roughly constant separation by recursive partitioning of triangles into 4 smaller triangles by connecting midpoints with a initial condition of the 20 triangles of an icosahedron.
- (b) The rotations around the direction can be selected to be uniformly located on the unit circle.
- (c) Combine the triangulation and circle abscissas to get a set of Euler angle abscissas for the projection orientation.

3. Apply standard software to the homogeneous reconstruction and the Euler angle abscissas in order to get a set of template images, one image for each projection direction abscissa. The scale of initial calculations is $N_T = 40 \times 10^3$ template images.

4. For the i th image in the stack:

- (a) Correlate the image with each template image. If the image in the stack is y_i and the j th template is z_j (both vectors) then the correlation is $r_{i,j} = y_i^T z_j$.

- (b) If the angular separation of the abscissas is roughly ν_a and the stated orientation of the image stack as determined by the biologist user is roughly ν_r then for each stack image i collect the set of indices $J_i \subset \{1, \dots, N_T\}$ of the $(\lceil \nu_r / \nu_a \rceil)^3$ templates having the largest correlations. These could be in quite disjoint regions of the Euler angle space for at least two reasons: the orientation of the j th image in the stack truly has a multi-modal pdf or the orientation is near the boundary of the fundamental domain sampled by triangulation and part of a unimodal pdf gets mapped by symmetry to a different region of the Euler angle space.
- (c) Define the *a priori* pdf (really pmf) on the orientation of the i th image by

$$p_i''(j) = \begin{cases} |r(i, j)|, & j \in J_i \\ 0, & \text{otherwise} \end{cases} \quad (5.11)$$

$$p_i'(j) = \frac{p_i''(j)}{\sum_{j \in J_i} p_i''(j)} \quad (5.12)$$

$$p_i(j) = p_i'(j) \sin(\beta_j) \quad (5.13)$$

where β_j is the second Euler angle of the projection orientation of the j th template image.

- (d) For each template image j collect the set of indices $I_j \subset \{1, \dots, N_v\}$ of the images for which $p_i(j) > 0$. In Matlab [40] it is probably desirable to store I_j as $\mathbb{I}\{j\}$ ($\dot{\mathbb{I}}$) so that $\mathbb{I}\{j\}$ is a vector of indices that can be used to select columns out of the \mathbb{Y} matrix that stores the i th image as a vector in the i th column. For different values of j , there will be a different number of components in the vector $\mathbb{I}\{j\}$. If I_j is stored as a $Z^{N_T \times n}$ matrix then n must be large enough to contain the longest list and a second vector $h \in Z^{N_T}$ will be needed where h_j is the number of

elements in I_j . With this data structure, at a later point in the program it is possible to compute $L^T * y(:, I(j, 1 : h(j)))$.

- (e) Compute $u \in Z^{N_r}$ for which the j th component the number of elements in I_j .
- (f) In the reconstruction software the key part of the iterative computation is the following: Let j be the index of the `parfor` loop over all abscissas. Test if $u_j = 0$. If true then no work in the j th iteration of the `parfor` loop. If false then
 - i. Compute L for the j th set of Euler angles. This will likely imply the computation of more L values than are computed in the software used in Chapters 2 and 3.
 - ii. Compute $L^T * y(:, I\{j\})$. This will likely be for a small subset of all columns of y and therefore many fewer columns than are currently used in the software used in Chapters 2 and 3 and this is the desired saving of computational effort.

5.2 Software efficiency and supercomputing resources

The theory of earlier sections applies for any choice of basis functions. However, the software uses the specific basis functions described in Ref. [83] where each basis function is the product of an icosahedral harmonic and a spherical Bessel function. A software implementation of the method was written that is suitable for execution in either the proprietary Matlab [40] or open source Octave [47] engine on a shared-memory computer. The update of Q is done by `fminbnd` in both Matlab and Octave. The update of V is implemented only for

the case where V is a diagonal matrix and is done by `fmincon` in Matlab and `SQP` in Octave. The limits of the software are partly memory requirements and partly use of simple numerical linear algebra algorithms. As an example of algorithmic limitations, Eq. 2.33 is solved by LU decomposition where the F matrix (Eq. 3.25) and g vector (Eq. 3.26) are each computed without taking advantage of the fact that orientations that are close to each other lead to similar contributions to the integrals in Eqs. 3.25 and 3.26. Relative to memory requirements, in order to make efficient Matlab code, the data is treated as matrix with dimensions that are the number of pixels per image by the number of images. This is the largest data structure in the runs described in this chapter. With the exception of the results for HK97 which are described in Section 3.2.2, all of the results described in this thesis are based on running the software using the Matlab engine on a dual-cpu quad-core Xeon (E5430 at 2.66GHz) with 16GB memory. In order to fit a computation into this hardware-software system, using more images implies using fewer basis functions or visa versa. For the results described in Chapters 2 and 3, all calculations used 1200 images and 720 basis functions (the so-called Step 7 of Ref. [83]) and each reconstruction takes approximately 2 days.

The software has been ported to the Gordon XSEDE cluster at the San Diego Supercomputer Center (SDSC) which has Matlab [40]. As is described in Section 3.2.2, test cases have been run using the Matlab Distributed Computing Server (MDCS).

In order to make initial measurements of the potential performance improvement achievable by using parallel computation at SDSC, a test case was constructed which performs a calculation that is similar to the calculation described in Chapter 2 but much smaller. The software performance results are shown in

Figure 5.1 and Figure 5.2. The running time of the test case is much smaller than the running time of the biological problems described in Chapter 3, but using the small test cases allowed the examination of a much wider range of parameters in the computing environment.

Figure 5.1 shows the running time improvement achieved by increasing the number of parallel cores using both a local scheduler on a single node and a TORQUE resource manager (or scheduler) over cores from multiple nodes. The local scheduler is able to schedule both shared memory and distributed memory parallel jobs up to 12 cores. The TORQUE scheduler along with the Matlab Distributed Computing Server (MDCS) is able to run parallel jobs up to 128 cores. Although the local scheduler performs better in general than the TORQUE scheduler when the number of cores is below 12, the TORQUE scheduler and MDCS is able to achieve further speed-up of the calculation. In order to understand how close these performance curves are to linear speed-up, especially MDCS with 8 or more cores, Figure 5.2 plots the actual running time versus the number of cores in log-log scale (log with base 10) and the running time that would be achieved with linear speed-up. The achieved speed-up is less than linear for large numbers of cores. However, increasing the size of the problem to a biologically-relevant size will probably make the speed-up closer to linear because it will increase the ratio of computation to communication, which is artificially low in this test case.

In addition, system software at the San Diego Supercomputer Center is available that does virtual shared memory processing (vSMP) where multiple distributed-memory nodes of the cluster are made to appear to the user as a single node. This provides an alternative environment for running Matlab which

Time versus Cores

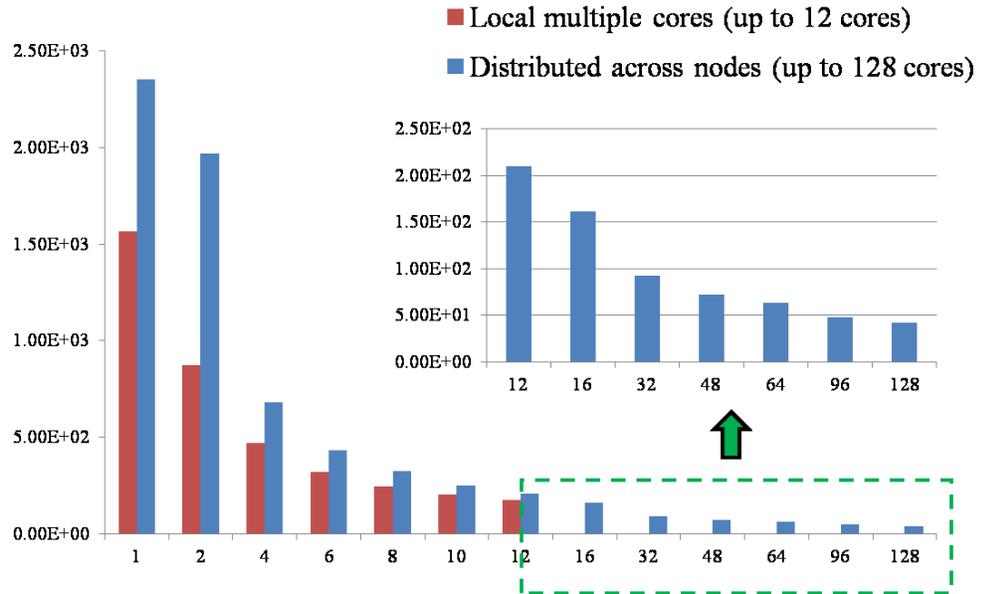


Figure 5.1: Running time from a test case versus number of cores. The red bars are from parallel jobs using the local scheduler. The blue bars are from parallel jobs using the Matlab Distributed Computing Server (MDCS) and the TORQUE scheduler of the San Diego Supercomputer Center (SDSC).

merits investigation.

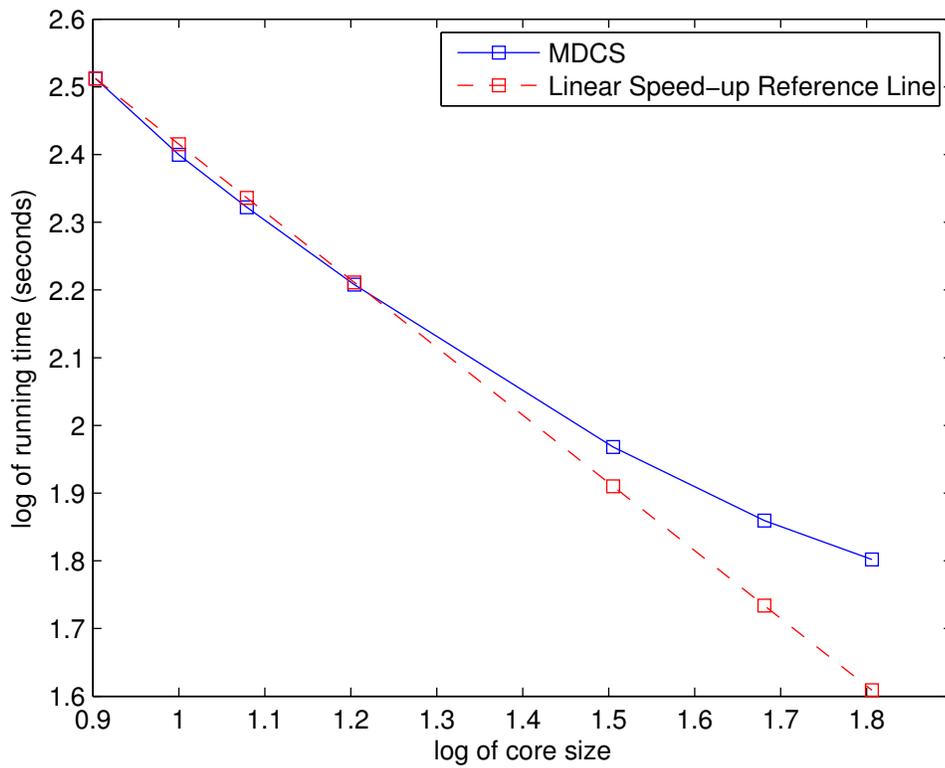


Figure 5.2: Running time from a test case versus the number of cores in log-log scale, and compare with the reference line of the linear speed-up.

CHAPTER 6

CONCLUSION

6.1 Conclusions of this thesis

The primary conclusions of this thesis are that maximum likelihood estimators that solve the heterogeneous particle reconstruction problem can be posed, solved, and implemented in practical software and that the resulting heterogeneous reconstructions are relevant to biology. In one sense the estimators are substantial generalizations of previous homogeneous reconstruction estimators [16, 34, 57, 83] because the new estimators treat the problem as a stochastic signal in noise problem rather than a deterministic signal in noise problem. In a second sense the estimators are substantial generalizations of previous pattern analysis estimators [8, 42, 53] because the new estimators do not measure the random vector with a Gaussian mixture pdf directly but rather through a stochastic linear transformation with the addition of noise. The relevance of the results to biology is most clearly demonstrated in the work on the maturation of $N\omega V$ which is described in Chapter 3.

6.2 Future directions and challenges

Sparse full-matrix covariance for the coefficients by L_1 optimization

The estimation of large matrices from limited data is an important current topic in statistics [6,7,11,31,43,46,52,54]. Methods for covariance matrices include [11] banding, tapering, thresholding, penalties, and regularization. An appropriate method for V_η depends in part on the choice of basis functions. For instance, using the voxel basis functions, every element (i, j) of V_η corresponds to a pair of triple-integer indices (\mathbf{m}, \mathbf{n}) for the corresponding voxels and it is natural that the elements of V_η decay as a function of $\|\mathbf{m} - \mathbf{n}\|_2$ where $\|\cdot\|_2$ is the Euclidean norm. On the other hand, using harmonic basis functions also leads to triple-integer indices but now two of the indices describe the linear combination of spherical harmonics and one describes the radial function and so $\|\mathbf{m} - \mathbf{n}\|_2$ no longer has a geometric interpretation. Possibly the most promising approach is to focus on regularizers [7, 52], transform the regularizer into an *a priori* pdf on V_η (e.g., square of the Euclidean norm of a vector transforms into a Gaussian *a priori* pdf), replace the maximum likelihood criteria by a maximum *a posteriori* (MAP) criteria, and continue to use expectation maximization to compute the solution. This program has been followed successfully [16, Section VII] for homogeneous reconstructions. If $\Theta = V_\eta^{-1}$ then a basic choice of regularizer [52] is $\|\Theta\|_{1,\text{off}} = \sum_{i \neq j} |\Theta_{i,j}|$. Cross validation is the standard method for determining the weight on the regularization term, though the computation burden may become a problem in larger applications. As is described in Section 5.1.3, it is anticipated that many users of these estimators will first solve a homogeneous object case,

then a heterogeneous case with a diagonal V_η (as is done in Chapters 2 and 3), and finally a heterogeneous case with general V_η . Using the answer from the previous step as the initial condition in the current step may allow efficient solution even if the overall MAP problem lacks properties such as convexity. Currently we use a trust region method (Matlab [40], `fmincon`) with analytical first and second derivatives with such initial conditions.

Faster computation

In order to have an impact on the biological community, better software is required. Two major senses in which the software needs to be improved are described in the following two paragraphs.

It is important to change to a programming language and user interface that are familiar to the biological community. Unfortunately, Matlab [40] is not familiar to the biological community. However, python [71], which has many similarities to when an appropriate set of libraries are used, is familiar to the community. Therefore a python implementation, which can include MPI [72–74] type message-passing parallel functionality, is attractive. Such an implementation could be integrated into the Appion system [32, 75] which would immediately provide a familiar user interface.

A second sense in which the software needs to be improved is through better algorithms to implement the estimators. Several areas of current effort are described in Chapter 5. In particular, the use of *a priori* information as described in Section 5.1.3 seems promising.

BIBLIOGRAPHY

- [1] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, December 1972. Tenth Printing.
- [2] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1979.
- [3] Matthew L. Baker, Junjie Zhang, Steven J. Ludtke, and Wah Chiu. Cryo-EM of macromolecular assemblies at near-atomic resolution. *Nature Protocols*, 5(10):1697–1708, 2010.
- [4] T. S. Baker, N. H. Olson, and S. D. Fuller. Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and Molecular Biology Reviews*, 63(4):862–922, December 1999.
- [5] Asbjørn Berge and Anne H. Schistad Solberg. Structured Gaussian components for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sensing*, 44(11):3386–3396, November 2006.
- [6] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.
- [7] Peter J. Bickel and Bo Li. Regularization in statistics. *Test*, 15(2):271–344, 2006.
- [8] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, Department of Electrical Engineering and Computer Science, University of California at Berkeley, April 1998.
- [9] B. Bothner, D. Taylor, B. Jun, K. K. Lee, G. Siuzdak, C. P. Schultz, and J. E. Johnson. Maturation of a tetravirus capsid alters the dynamic properties and creates a metastable complex. *Virology*, 334:17–27, 2005.
- [10] Doryen Bubeck, David J. Filman, and James M. Hogle. Cryo-electron microscopy reconstruction of a poliovirus-receptor-membrane complex. *Nature Structural & Molecular Biology*, 12:615–618, July 2005.

- [11] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38(4):2118–2144, 2010.
- [12] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Wadsworth Group, Pacific Grove, CA, 2nd edition, 2002.
- [13] R. Holland Cheng, Vijay S. Reddy, Norman H. Olson, Andrew J. Fisher, Timothy S. Baker, and John E. Johnson. Functional implications of quasi-equivalence in a $T = 3$ icosahedral animal virus established by cryo-electron microscopy and x-ray crystallography. *Structure*, 2:271–282, 15 April 1994.
- [14] William Cochran, Francis H. C. Crick, and Vladimir Vand. The structure of synthetic polypeptides. I. The transform of atoms on a helix. *Acta Cryst.*, 5:581–586, 1952.
- [15] Satya Dharanipragada and Karthik Visweswariah. Gaussian mixture models with covariances or precisions in shared multiple subspaces. *IEEE Trans. Audio, Speech, and Language Proc.*, 14(4):1255–1266, July 2006.
- [16] Peter C. Doerschuk and John E. Johnson. *Ab initio* reconstruction and experimental design for cryo electron microscopy. *IEEE Transactions on Information Theory*, 46(5):1714–1729, August 2000.
- [17] Tatiana Domitrovic, Tsutomu Matsui, and John E. Johnson. Dissecting quasi-equivalence in non-enveloped viruses: Membrane disruption is promoted by lytic peptides released from subunit pentamers, not hexamers. *J. Virology*, 2012. VI Accepts, published online ahead of print on 3 July 2012. *J. Virol.* doi:10.1128/JVI.01089-12.
- [18] Bradley Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–487, December 1978.
- [19] Harold P. Erickson. The Fourier transform of an electron micrograph—First order and second order theory of image formation. In R. Barer and V. E. Cosslett, editors, *Advances in Optical and Electron Microscopy (Volume 5)*, pages 163–199. Academic Press, London and New York, 1973.
- [20] Andrew J. Fisher and John E. Johnson. Ordered duplex RNA controls capsid architecture in an icosahedral animal virus. *Nature*, 361:176–179, Jan. 14 1993.

- [21] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego, 1996.
- [22] S. D. Fuller, S. J. Butcher, R. H. Cheng, and T. S. Baker. Three-dimensional reconstruction of icosahedral particles—the uncommon line. *J. Struct. Biol.*, 116(1):48–55, January 1996.
- [23] George Harauz and Marin van Heel. Exact filters for general geometry three dimensional reconstruction. *Optik*, 73(4):146–156, 1986.
- [24] Charlotte Helgstrand, Sanjeev Munshi, John E. Johnson, and Lars Liljas. The refined structure of *Nudaurelia capensis* ω Virus reveals control elements for a $T = 4$ capsid maturation. *Virology*, 318:192–203, 2004.
- [25] Roger W. Hendrix and John E. Johnson. Bacteriophage HK97 capsid assembly and maturation. *Adv. Exp. Med. Biol.*, 726:351–363, 2012. doi: 10.1007/978-1-4614-0980-9_15.
- [26] James M. Hogle. Poliovirus cell entry: Common structural themes in viral cell entry pathways. *Annu. Rev. Microbiol.*, 56:677–702, 15 July 2002.
- [27] Rick K. Huang, Reza Khayat, Kelly K. Lee, Ilya Gertsman, Robert L. Duda, Roger W. Hendrix, and John E. Johnson. The Prohead-I structure of bacteriophage HK97: Implications for scaffold-mediated control of particle assembly and maturation. *J. Molecular Biology*, 408(3):541–554, 6 May 2011. <http://dx.doi.org/10.1016/j.jmb.2011.01.016>.
- [28] Grant J. Jensen, editor. *Cryo-EM, Part A: Sample Preparation and Data Collection*, volume 481 of *Methods in Enzymology*. Elsevier Inc., 2010.
- [29] Grant J. Jensen, editor. *Cryo-EM, Part B: 3-D Reconstruction*, volume 482 of *Methods in Enzymology*. Elsevier Inc., 2010.
- [30] Grant J. Jensen, editor. *Cryo-EM, Part C: Analyses, Interpretation, and Case studies*, volume 483 of *Methods in Enzymology*. Elsevier Inc., 2010.
- [31] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254–4278, 2009.
- [32] Gabriel C. Lander, Scott M. Stagg, Neil R. Voss, Anchi Cheng, Denis Fellmann, James Pulokas, Craig Yoshioka, Christopher Irving, Anke Mulder,

- Pick-Wei Lau, Dmitry Lyumkis, Clinton S. Potter, and Bridget Carragher. Appion: An integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.*, 166(1):95–102, 2009.
- [33] J. Lanman, J. Crum, T. J. Deerinck, G. M. Gaietta, A. Schneemann, G. E. Sosinsky, M. H. Ellisman, and J. E. Johnson. Visualizing flock house virus infection in *Drosophila* cells with correlated fluorescence and electron microscopy. *J. Struct. Biol.*, 161:439–446, 2008.
- [34] Junghoon Lee, Peter C. Doerschuk, and John E. Johnson. Exact reduced-complexity maximum likelihood reconstruction of multiple 3-D objects from unlabeled unoriented 2-D projections and electron microscopy of viruses. *IEEE Transactions on Image Processing*, 16(11):2865–2878, November 2007.
- [35] Seunghee Lee, Peter C. Doerschuk, and John E. Johnson. Reciprocal space representation of helical objects and their projection images for helices constructed from motifs without spherical symmetry. *Ultramicroscopy*, 109:253–263, 2009.
- [36] Seunghee Lee, Peter C. Doerschuk, and John E. Johnson. Multi-class maximum likelihood symmetry determination and motif reconstruction of 3-D helical objects from projection images for electron microscopy. *IEEE Transactions on Image Processing*, 20(7):1962–1976, July 2011.
- [37] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer-Verlag, New York, 2 edition, 1998.
- [38] J. Lepault and T. Pitt. Projected structure of unstained, frozen-hydrated T-layer of *bacillus brevis*. *EMBO J.*, 3(1):101–105, 1984.
- [39] Hongrong Liu, Lei Jin, Sok Boon S. Koh, Ivo Atanasov, Stan Schein, Lily Wu, and Z. Hong Zhou. Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science*, 329:1038–1043, 27 August 2010.
- [40] Mathworks URL. <http://www.mathworks.com/>.
- [41] Tsutomu Matsui, Gabriel C. Lander, Reza Khayat, and John E. Johnson. Subunits fold at position-dependent rates during maturation of a eukaryotic RNA virus. *Proc. Nat. Acad. Sci. U.S.A.*, 107(32):14111–14115, 10 Aug. 2010.

- [42] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 1997.
- [43] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.
- [44] M. F. Moody. Image analysis of electron micrographs. In P. W. Hawkes and U. Valdrè, editors, *Biophysical Electron Microscopy: Basic Concepts and Modern Techniques*, chapter 7, pages 145–287. Academic Press, 1990.
- [45] S. Munshi, L. Liljas, J. Cavarelli, W. Bomu, B. McKinney, V. Reddy, and J. E. Johnson. The 2.8 structure of a $T = 4$ animal virus and its implications for membrane translocation of RNA. *J. Mol. Biol.*, 261(1):1–10, 9 August 1996.
- [46] Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.
- [47] Octave URL. <http://www.gnu.org/software/octave/>, <http://octave.sourceforge.net/>.
- [48] Pawel A. Penczek, Chao Yang, Joachim Frank, and Christian M. T. Spahn. Estimation of variance in single-particle reconstruction using the bootstrap technique. *J. Struct. Biol.*, 154(2):168–183, 2006.
- [49] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–1612, 2004.
- [50] Stephen W. Provencher and Robert H. Vogel. Three-dimensional reconstructions from electron micrographs of disordered specimens: I. Method. *Ultramicroscopy*, 25:209–222, 1988.
- [51] Cory J. Prust, Peter C. Doerschuk, Gabriel C. Lander, and John E. Johnson. *Ab initio* maximum likelihood reconstruction from cryo electron microscopy images of an infectious virion of the tailed bacteriophage P22 and maximum likelihood versions of Fourier Shell Correlation appropriate for measuring resolution of spherical or cylindrical objects. *J. Struct. Biol.*, 167:185–199, 2009.
- [52] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l_1 -penalized log-

- determinant divergence. *ArXiv:0811.3628v1 [stat.ML]*, pages 1–35, 21 Nov. 2008.
- [53] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.
- [54] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.
- [55] Sjors H. W. Scheres. Classification of structural heterogeneity by maximum-likelihood methods. In Grant J. Jensen, editor, *Cryo-EM, Part B: 3-D Reconstruction*, volume 482, pages 295–320. Elsevier, Inc., 2010.
- [56] Sjors H. W. Scheres. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.*, 415(2):406–418, 13 January 2012.
- [57] Sjors H. W. Scheres, Haixiao Gao, Mikel Valle, Gabor T. Herman, Paul P. B. Eggermont, Joachim Frank, and Jose-Maria Carazo. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1):27–29, January 2007.
- [58] F. J. Sigworth. A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.*, 122:328–339, 1998.
- [59] Fred J. Sigworth, Peter C. Doerschuk, Jose-Maria Carazo, and Sjors H. W. Scheres. An introduction to maximum-likelihood methods in cryo-EM. In Grant J. Jensen, editor, *Cryo-EM, Part B: 3-D Reconstruction*, volume 482, pages 263–294. Elsevier, Inc., 2010.
- [60] Khe Chai Sim and Mark J. F. Gales. Minimum phone error training of precision matrix models. *IEEE Trans. Audio, Speech, and Language Proc.*, 14(3):882–889, May 2006.
- [61] Hichem Snoussi and Ali Mohammad-Djafari. Estimation of structured Gaussian mixtures: The inverse EM algorithm. *IEEE Trans. Sig. Proc.*, 55(7):3185–3191, July 2007.
- [62] M. Stark, Harold. *An Introduction to Number Theory*. Markham Publishing Company, 1970.
- [63] Jinghua Tang, Jennifer M. Johnson, Kelly A. Dryden, Mark J. Young, Adam Zlotnick, and John E. Johnson. The role of subunit hinges and molecular

- “switches” in the control of viral capsid polymorphism. *J. Struct. Biol.*, 154(1):59–67, April 2006.
- [64] Jinghua Tang, Kelly K. Lee, Brian Bothner, Timothy S. Baker, Mark Yeager, and John E. Johnson. Dynamics and stability in maturation of a $t = 4$ virus. *J. Molecular Biology*, 392:803–812, 2009.
- [65] Jinghua Tang, Qiu Wang, Brad M. Kearne, Peter C. Doerschuk, Timothy S. Baker, and John E. Johnson. In preparation. 2013.
- [66] D. J. Taylor, N. K. Krishna, M. A. Canady, A. Schneemann, and J. E. Johnson. Large-scale, pH-dependent, quaternary structure changes in an RNA virus capsid are reversible in the absence of subunit autoproteolysis. *J. Virology*, 76:9972–9980, 2002.
- [67] Ye Tian, Jian-Lai Zhou, Hui Lin, and Hui Jiang. Tree-based covariance modeling of hidden Markov models. *IEEE Trans. Audio, Speech, and Language Proc.*, 14(6):2134–2146, November 2006.
- [68] Mariana Tihova, Kelly A. Dryden, Thucvy L. Le, Stephen C. Harvey, John E. Johnson, Mark Yeager, and Anette Schneemann. Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length. *J. Virol.*, 78(6):2897–2905, 2004.
- [69] Chikashi Toyoshima and Nigel Unwin. Contrast transfer for frozen-hydrated specimens: Determination from pairs of defocused images. *Ultramicroscopy*, 25(4):279–291, 1988.
- [70] URL. <http://www.emdatbank.org/>.
- [71] URL. <http://www.python.org/>.
- [72] URL. <http://www.mpi-forum.org/>.
- [73] URL. <http://www.mcs.anl.gov/research/projects/mpi/>.
- [74] URL. <http://www.open-mpi.org/>.
- [75] URL. <http://ami.scripps.edu/redmine/projects/appion/wiki>.

- [76] URL. Flock House Virus (FHV) web page. http://viperdb.scripps.edu/info_page.php?VDB=2q25.
- [77] VIPERdb URL. <http://viperdb.scripps.edu/>.
- [78] Marin van Heel. Similarity measures between images. *Ultramicroscopy*, 21:95–100, 1987.
- [79] R. H. Vogel, S. W. Provencher, C.-H. von Bonsdorff, M. Adrian, and J. Dubochet. Envelope structure of Semliki Forest virus reconstructed from cryo-electron micrographs. *Nature*, 320:533–535, 10 April 1986.
- [80] Robert H. Vogel and Stephen W. Provencher. Three-dimensional reconstructions from electron micrographs of disordered specimens: II. Implementation and results. *Ultramicroscopy*, 25:223–240, 1988.
- [81] Qiu Wang. *Maximum likelihood reconstruction of heterogeneous 3-D objects from 2-D projections of unknown orientation and application to electron microscope images of viruses*. PhD thesis, School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, USA, June 2013.
- [82] Qiu Wang, Tsutomu Matsui, Tatiana Domitrovic, Yili Zheng, Peter C. Doerschuk, and John E. Johnson. Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps. *J. Struct. Biol.*, 181(3):195–206, March 2013. <http://dx.doi.org/10.1016/j.jsb.2012.11.005>.
- [83] Zhye Yin, Yili Zheng, Peter C. Doerschuk, Padmaja Natarajan, and John E. Johnson. A statistical approach to computer processing of cryo electron microscope images: Virion classification and 3-D reconstruction. *J. Struct. Biol.*, 144(1/2):24–50, 2003.
- [84] Xing Zhang, Ethan Settembre, Chen Wu, Philip R. Dormitzer, Richard Bellamy, Stephen C. Harrison, and Nikolaus Grigorieff. Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc. Nat. Acad. Sci. U.S.A.*, 105(6):1867–1872, 12 February 2008.
- [85] Yibin Zheng and Peter C. Doerschuk. 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.

- [86] Yili Zheng. *Novel statistical models and a high-performance computing toolkit for the solution of cryo electron microscopy inverse problems in viral structural biology*. PhD thesis, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA, August 2008.
- [87] Yili Zheng, Qiu Wang, and Peter C. Doerschuk. 3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object. *Journal of the Optical Society of America A*, 29(6):959–970, June 2012. Zheng and Wang are co-first authors. <http://dx.doi.org/10.1364/JOSAA.29.000959>.