Algorithms for Image Restoration and 3D Reconstruction from Cryo-EM Images

Tejal Bhamre

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Physics
Advisers: Amit Singer and Joshua Shaevitz

June 2017

© Copyright by Tejal Bhamre, 2017.

All Rights Reserved

Abstract

Single particle reconstruction (SPR) in cryo-electron microscopy (cryo-EM) has recently emerged as the method of choice to determine the structure of biological macromolecules to near atomic resolution. The typical procedure for obtaining the final high resolution 3D structure is by starting with an initial guess and iteratively refining it using the acquired dataset of the molecule's 2D projection images. The final estimate from the refinement procedure is known to often depend heavily on the initial model used as the starting point, thereby making a good initial estimate crucial for success.

In this thesis, we propose and test two novel approaches, which we call Orthogonal Extension and Orthogonal Replacement, for 3D ab-initio and homology modeling in SPR using cryo-EM and X-ray free electron lasers (XFEL). Our approach is inspired by the molecular replacement technique used in X-ray crystallography. We first test both approaches on noisy synthetic datasets.

Motivated by the need for a reliable estimator of the covariance matrix, we develop a new image restoration method to perform contrast transfer function (CTF) correction and denoising in a single step. Through results on several experimental datasets, we demonstrate the efficacy of our method as a single, preliminary step to inspect particle images, detect outliers, and estimate the covariance matrix of the underlying clean images. Our covariance matrix estimator is asymptotically consistent and successfully corrects for the CTF.

An immediate application of improved covariance estimation is an improvement in the 2D classification or class averaging procedure in the cryo-EM pipeline. We digress from 3D homology/ab-initio modeling to focus on this application. Since different cryo-EM images are affected by noise as well as different CTF's or point spread functions from the microscope, the Euclidean distance between two images is not an optimal metric for their affinity. We derive and test a new affinity measure

akin to the Mahalanobis distance to compare cryo-EM images belonging to different defocus groups. We demonstrate that the new metric leads to an improvement in nearest neighbor detection and therefore the obtained class averages.

Finally, we revisit the homology modeling procedure of Orthogonal Extension. We incorporate our improved covariance matrix estimator into the Orthogonal Extension algorithm and propose a family of asymptotically unbiased estimators to recover the 3D structure. We demonstrate the advantage of our estimator through numerical experiments on synthetic and experimental datasets. We foresee this method as a good way to provide models to initialize refinement, directly from experimental images without performing class averaging and orientation estimation in cryo-EM and XFEL. Our second algorithm for ab-initio modeling, Orthogonal Replacement, is tested on synthetic datasets. In future work, Orthogonal Replacement would require designing an appropriate experiment to collect datasets that would facilitate its usage.

Publications and Presentations associated with this Dissertation

The material in this dissertation has appeared in the following publications. Most chapters have been slightly modified from the published articles. Nevertheless, the copyright to the original articles rests with the relevant journals.

1. *Orthogonal matrix retrieval in cryo-electron microscopy*, Tejal Bhamre, Teng Zhang, and Amit Singer, 12th IEEE International Symposium on Biomedical Imaging, pp. 1048–1052 (2015).

2. *Denoising and Covariance Estimation of Single Particle Cryo-EM Images*, Tejal Bhamre, Teng Zhang, and Amit Singer, Journal of Structural Biology, 195 (1), pp. 72-81 (2016).

3. *Mahalanobis Distance for Class Averaging of Cryo-EM Image*, Tejal Bhamre, Zhizhen Zhao, and Amit Singer, 14th IEEE International Symposium on Biomedical Imaging (2017), accepted.

4. *Anisotropic Twicing for Single Particle Reconstruction using Autocorrelation Analysis*, Tejal Bhamre, Teng Zhang, and Amit Singer, submitted.

The following public presentations were based on the materials in this dissertation:

1. Invited talk by Teng Zhang at Imaging and Modeling in Electron Microscopy - Recent Advances, Banff International Research Station, May 2014.

2. Poster at Semidefinite Programming and Graph Algorithms, Institute for Computational and Experimental Research in Mathematics (ICERM), February 2015.

3. Seminar at the Program in Applied and Computational Mathematics, Princeton University, February 2015.

4. Poster at the 12th IEEE International Symposium on Biomedical Imaging, April 2015.

5. Invited talk at SIAM Conference on Imaging Science, May 2016.

6. Invited talk at Workshop on Computational Methods for Cryo-EM, Simons Electron Microscopy Center, October 2016.

Acknowledgements

Working towards a PhD is an incredibly humbling and transformative journey. There are many people who have made this journey possible and rewarding for me. I'm extremely grateful to my adviser, Amit Singer, for being a great mentor and for his patience, wisdom, and enthusiasm. When I made the somewhat late transition from theoretical physics to applied mathematics in the third year of graduate school, Amit welcomed me to the many interesting challenges in the world of cryo-EM. I've learnt a great deal from our technical discussions and the work in this thesis would have been impossible without Amit's valuable insights and suggestions. I thank him for allowing me enough freedom and time to solve problems, while ensuring that I did not go astray. I thank my physics advisor, Joshua Shaevitz, for his interest and enthusiasm about my work. I'd like to acknowledge my committee members and readers for their valuable inputs and comments. I'm deeply indebted to Paul Steinhardt and Salvatore Torquato. My journey in graduate school began with dabbling in condensed matter physics problems under their guidance. I cherish all the interesting problems concerning hyperuniformity and quasicrystals that I learnt from them. I'm grateful to Paul for encouraging me to recognize and pursue a different research direction in cryo-EM.

I had a wonderful experience working in William Happer's atomic physics lab as part of my experimental qualifying exams. I thank my colleagues in Will's group, Iannis, Bart, and Robert, for helping the theorist me figure out how to work on experiments in the laboratory. I would like to thank my collaborators, Teng Zhang and Jane Zhao, from whom I have learnt a great deal. Many thanks to Fred Sigworth and his group for organizing a tour of the electron microscopes at Yale, and his many helpful suggestions about my work. I'm grateful to Adam Frost, and his students, for giving me an inside look into the electron microscopes at UCSF. I'm very thankful for collaborations and enlightening discussions with Yoel Shkolnisky, Joakim Anden,

Xiuyuan Cheng, and Lanhui Wang. I'm grateful to Herman Verlinde, the physics department chair, for helpful discussions about navigating through grad school.

I would like to express my deepest gratitude to all my teachers, friends, and colleagues throughout school and undergraduate studies - it hard to do justice in this limited space to the role that all of them have played in shaping my thoughts. My first encounter with research in theoretical physics was through a summer project with Urjit Yajnik, which resumed later for my senior thesis. I thoroughly enjoyed our conversations on cosmology and physics in general. It is hard to exaggerate the role that Raghava Varma and Achim Kempf have played in my decision to pursue a PhD. Raghava Varma motivated me tremendously to pursue research in his role as my junior thesis advisor, and made a prophetic comment in passing that I would enjoy research in applied math better, which I came to realize for myself only much later in grad school. I worked as a research assistant in Kempf's lab at the University of Waterloo, Canada. This led to a very fruitful and enjoyable collaboration on a quantum gravity problem, and many frequent trips to Waterloo followed. I have had the pleasure of many fascinating conversations with him about physics, history, Bollywood movies, and traveling the world! My sincere thanks to him and Dushyantha for being such fun and gracious hosts, and sightseeing Montreal with me.

Graduate school is not just about research. I'm indebted to Shefalika Gandhi, whose mindfulness and meditation workshops in Princeton have been a steady source of comfort and relief. I feel lucky to have been surrounded by many wonderful friends who have made my time at Princeton memorable - Chinmay Khandekar, Debajit Bhattacharya, Jahnavi Punekar, Jaya Khanna, Nayana Prasad, Paula Mateo, Pathikrit Bhattacharya, Ravi Tandon, Sneha Rath, Srinivas Narayana (NG), Tanya Gupta, Yogesh Goyal. I feel blessed to have had lovely roommates who are 'almost-sisters' to me now - thank you Jahnavi Punekar and Nayana Prasad, for being there every single day. I cherish all the fun we had, the food we cooked and devoured

tirelessly, and also all the hours we spent complaining about and over-analyzing grad school. Special thanks to Srinivas for giving me company when we worked on our respective deadlines, and for tolerating my never ending philosophical questions. I've enjoyed many musical evenings singing with Srinivas, Debajit, Jahnavi, with our other friends as the audience (sometimes forcefully). Jaya and Pathikrit have been a great source of comfort and fun - thank you for all the dinners, lunches, and music. I miss all the delightful girls' nights with Jahnavi and Paula, and Nayana, Sneha and Tanya. Yogesh and Debajit patiently trained me to shoot three pointers on the basketball court. Thanks to Yogesh for all the fights and annoying me to no end but always making up for it with a cup of the most perfect chai. Discussions with Chinmay, NG, Ravi about philosophy, life, and everything else were (and are) always intense and stimulating. Skype sessions and chats with Anasuya Mandal brightened my days and always made me smile. Thanks to Yash Deshpande for sharing his office space with me at Stanford, and for all the uplifting dog pictures. I spent a summer in Seattle which Arnab and Shaoni Sinha made memorable by their company and hospitality.

I have enjoyed the company of everyone in Amit's group - lunches with my groupmates Amit, Joao, Jose, Susannah, Yuehaw, Yuan, Yutong, provided the much needed break from research in the office. I'm thankful to all the wonderful staff of PACM and the physics department who have been of great help - Angela Lewis, Audrey Mainzer, Ben Rose, Charlene Borsack, Darryl Johnson, Jessica Heslin, Kate Brosowsky, Laura Deevey, Lisa Giblin, Vinod Gupta.

A special thanks to my favorite coffee shops (and the physics and math coffee lounge) in Princeton and Mountain View that have powered the work in this thesis.

None of this would have been possible without the support of my family. My late grandmother, Snehalata Padalkar, was my constant companion and biggest champion growing up. This thesis is dedicated to her boundless faith and love. My parents, Santosh and Bharati Bhamre, have relentlessly supported and encouraged me to pursue

my dreams. I cannot possibly convey my gratitude towards them in words. I thank my in-laws for their love. My biggest pillar of strength throughout PhD has been my best friend and now husband, Saket, who has been there with me through it all, despite the 3000 miles between us. It's impossible to put into words what he means to me, but thankfully I think he already knows. Thank you for braving through the deadlines, stressful months and the accompanying crankiness, walking with me every step of the way, taking over the chores, cheering me up, and most importantly, never letting me lose sight of the bigger, wonderful landscape of life, with you.

To my grandmother.

List of Acronyms

CCD	Charge Coupled Device
Cryo-EM	Cryo-electron Microscopy
CTF	Contrast Transfer Function
FCR	Fourier Cross Resolution
FSC	Fourier Shell Correlation
MR	Molecular Replacement
PCA	Principal Component Analysis
SDP	Semidefinite Programming
SNR	Signal to Noise Ratio
SSNR	Spectral Signal to Noise Ratio
SPR	Single Particle Reconstruction
SVD	Singular Value Decomposition
VDM	Vector Diffusion Maps
XFEL	X-ray Free Electron Microscopy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."
>
> Watson and Crick, Nature [107]

"Structure is function" is an often heard mandate in structural biology. In 1962, the Nobel Prize in medicine was awarded to Crick, Watson and Wilkins for discovering the double helix structure of DNA. The detailed knowledge of a biological macromolecules's structure, including the 3D arrangement of its atoms, is the first step in understanding its function and hence, biological mechanisms.

Since then, hundreds of thousands of macromolecular structures have been solved. The popular method of choice to solve macromolecular structures for several decades has been X-ray crystallography, in which the diffraction patterns of X-rays scattered from a crystallized protein are used to deduce its structure. Crystallization is required to amplify the weak signal, that is, the diffraction pattern of a single molecule. A large, perfectly ordered crystal can produce a diffraction pattern that is resolvable to a sufficiently high resolution to allow 3D reconstruction. However, a serious

limitation to X-ray crystallography is that many proteins and viruses are resistant to crystallization, thereby limiting the ability to study these macromolecules. Enter single particle cryo-electron microscopy (cryo-EM). Although cryo-EM was first introduced as early as the 1970's, it has grown into a "revolution" resisting crystallization [44] since 2013, with its rapid advancement in leaps and bounds. Cryo-EM allows the study of macromolecules in vivo, in their functionally active state, unlike X-ray crystallography, in which the process of crystallization may change the natural conformation of the complex.

In Single Particle Reconstruction (SPR) using cryo-EM [23, 44, 5, 60], the sample of macromolecules is rapidly frozen in a thin vitreous ice layer, maintained at liquid nitrogen temperature, and imaged with an electron microscope to acquire two dimensional projection images of the macromolecule at random, unknown directions. The specimen, consisting of an ensemble of macromolecules, is imaged to acquire several large top views called 'micrographs' from which individual particle images are subsequently selected. The main advantage of cryo-EM over X-ray crystallography is that the specimen can be studied in its native state and does not need to be crystallized. Moreover, cryo-EM can be used to study molecules that exhibit structural variability and conformational changes. Additionally, unlike X-ray crystallography, cryo-EM requires only small amounts of sample, making it possible to obtain structures from samples that cannot be isolated in large enough quantities for X-ray crystallography. Recent technological advances have resulted in near-atomic resolution 3D structures of complexes such as viruses, ribosomes, ion channels, as small as 170 kDa. For scale, a Dalton is 1/12'th the mass of a carbon atom.

SPR with cryo-EM is a very challenging problem due to multiple particularly difficult obstacles, mainly: the extremely low signal to noise ratio of the acquired images (see Fig. 1.1), and the problem of estimating the unknown orientations of the 2D projection images. The imaging process with electrons leads to radiation

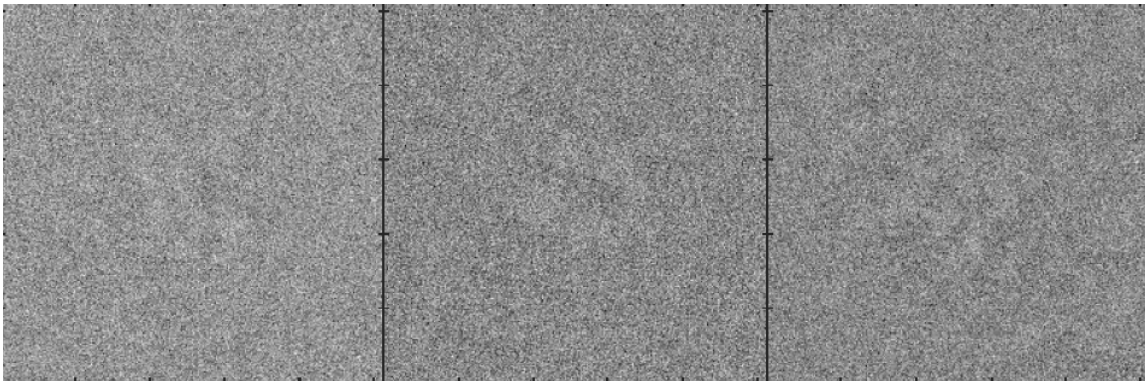


Figure 1.1: Three raw cryo-EM images from an experimental dataset of TRPV1 [47]

damage of the specimen, therefore one is constrained to low electron doses resulting in high levels of noise. The typical voltage used in electron microscopes ranges from 200 to 300kV. Additionally, determining the unknown orientations of the images is a challenging and computationally demanding non linear optimization problem.

1.1 The Cryo-EM Revolution

In 2015, cryo-EM was selected as the "Method of the Year" by the journal Nature Methods [2]. In the last five years, the Electron Microscopy Data Bank (EMDB) has been inundated with the deposition of high resolution structures. This wave of progress has been made possible by improved detector technology on one hand, along with better automation and software on the other.

The "goodness" of a detector is captured by its detective quantum efficiency (DQE), which quantifies how much the signal to noise ratio (SNR) of the measurement is affected by the process of detection [17]. The DQE ranges from 0 to 1, with 1 being characteristic of an ideal detector. Until recently, charge-coupled device (CCD) cameras were preferred for cryo-EM despite having a low DQE ~ 0.1 at high energies, due to the ease of automated data acquisition with them [77]. In CCD cameras, the incident electrons are first converted into photons via a scintillator, a mechanism that

is lossy and leads to further degradation of the already weak signal. Starting around 2012, a new generation of 'direct electron detectors' spurred the onset of the current cryo-EM revolution [21]. Direct electron detectors, as the name suggests, are able to detect electrons directly without first converting them to photons, resulting in a much higher DQE than CCD cameras. They also provide much faster readout, enabling collection of cryo-EM data in the 'movie mode' in which several images are recorded in rapid succession. Electrons scattering from the cryo-EM sample cause movements of the sample leading to motion blur. With the captured movies, beam induced motion can now be tracked and corrected, facilitating de-blurring and retention of higher resolution information in images than before. Niko Grigorieff and his colleagues first demonstrated beam induced motion correction by leveraging the movie mode.

The improvement in detector technology was accompanied by new maximum likelihood based image analysis approaches, introduced by Fred Sigworth [85] in 1998 [84]. The first notable results that heralded the cryo-EM revolution were from the University of California in San Francisco (UCSF), where a $3.3\mathring{A}$ resolution map of the 20S proteasome was obtained using the K2 detector,and the Medical Research Council (MRC) in Cambridge, where ribosome structure details up to $4\mathring{A}$ resolution were obtained from merely $35,000$ particle images from a Falcon detector [45, 46, 13]. This was a significant milestone for cryo-EM from the $40\mathring{A}$ structures in the early "blog-ology" era of the 1990's.

1.2 Algorithmic Overview of Single Particle Reconstruction

The first step in SPR is collection of the data itself. The specimen frozen in an ice layer is imaged using an electron microscope. A top view of the sample is acquired in the form of a large image called a micrograph. The subsequent steps of the SPR

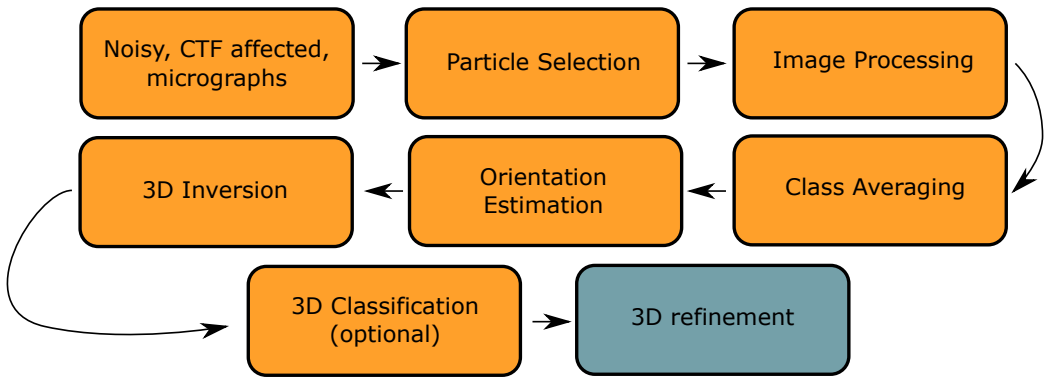


Figure 1.2: Cryo-EM Pipeline

pipeline are as follows (see Fig. 1.2):

- Particle Selection: Hundreds of thousands of individual particles need to be identified and picked from micrographs in the dataset, either semi-automatically or automatically. This step often needs manual intervention thereby making it time-consuming and prone to inconsistency [81, 61]. There have been many recent advancements to make particle selection fully automated, for instance, using deep learning and convolutional neural networks [104].

- Image Preprocessing: Raw cryo-EM images are extremely noisy and suffer from additional effects of the contrast transfer function (CTF) of the microscope. Raw images are first preprocessed by whitening, normalization, denoising and other image processing techniques to prepare them for the next step.

- Class Averaging or 2D Classification: Since cryo-EM images suffer from very low SNR's, neighboring images are first identified by partitioning the dataset using classification techniques. After alignment and averaging, the resulting images, called class averages, have a higher SNR and can be used for orientation estimation.

- Orientation Estimation: Next, the viewing angles or pose parameters associated with the images are estimated using angular reconstitution or common-line

based algorithms that leverage information in the common lines between pairs of images. This is possible due to the Fourier Projection Slice Theorem (see Appendix). Algorithms in the ASPIRE toolbox use convex optimization and spectral methods to achieve an orientation assignment that is as consistent as possible with all common lines.

- 3D classification (optional): Cryo-EM images are often obtained from a heterogeneous mixture of a macromolecule in two or more of its conformations. It is thus important to reconstruct all possible 3D conformations from the images. This is called the 'heterogeneity problem' in cryo-EM [37, 94].

- 3D Inversion: Using classical tomography techniques, the 3D structure is reconstructed from 2D images whose orientations have been assigned.

- Iterative Refinement: Using an ab-initio estimate as the starting 3D structure, the iterative refinement process uses the current estimate and the raw data to improve and refine the 3D structure until convergence.

1.2.1 Cryo-EM Software and Databases

The EMDB, an archive for 3D EM reconstructions, now contains over 4000 deposited maps, including several high resolution maps reflective of the "resolution revolution". In addition to final 3D structures, the need to publicly archive raw EM data for benchmarking and validation has been recognized by the cryo-EM community. The Electron Microscopy Public Image Archive (EMPIAR) is a repository dedicated to raw cryo-EM datasets. EMPIAR is designed to support large dataset transfer and contains terabytes of "big data" in the form of raw micrographs as well as movies.

There are many excellent open source software packages for structure reconstruction from cryo-EM images using statistical methods, such as RELION, SPIDER, Xmipp, FREALIGN, SPARX, EMAN2, SIMPLE, IMAGIC, etc. [96, 56, 82, 103,

80, 27]. The algorithms in this dissertation are included and in the software toolbox ASPIRE, publicly available at `spr.math.princeton.edu`.

SPR is a very challenging and computationally expensive inverse problem. The widespread accessibility of Graphical Processing Units (GPUs) in the recent years have vastly reduced the time needed to obtain 3D reconstructions from raw data [39].

1.3 3D Ab-Initio and Homology Modeling

The iterative refinement procedure in SPR requires an initial structure as the starting point. Starting from a good initial model can significantly reduce the number of iterations needed for convergence, and therefore lead to substantial savings in computational running time. Most refinement algorithms are based on some local optimization procedure, so there is no guarantee of converging to the global optimum solution. In particular, cryo-EM refinement suffers from the problem of 'model bias', meaning that the final solution is heavily influenced by the initial ab-initio model [85].

The quality of the initial model is less critical when the cryo-EM data itself enjoys a high SNR. In that case, relatively featureless initial models like ellipsoids can also converge to the global optimum using high quality data. However, in practice, the SNR is typically too low to allow this. For small complexes in particular, the images can be extremely noisy and refinement is challenging. There are a few approaches for determining a good ab-initio model to start the iterative refinement process. The method of moments and random conical tilt reconstruction provide low resolution ab-initio models [50, 51, 76, 99, 25] . However, the method of moments is very sensitive to errors in the data and hence not very reliable in practice.

Yet another approach is based on common lines between images, also known as angular reconstitution. This is based on the observation that the common lines between

three projection images uniquely determine their relative orientations up to chirality. There are several existing common lines based algorithms [68, 20, 55, 87, 83]. In [115], the authors obtained ab-initio models for the E. coli 50S ribosomal subunit from 27,121 projection images and for the 70S ribosome from 40,779 projection images. With these models, the refinement process converged in one or two steps. However, all these algorithms require computation of class averages to improve the SNR of the images enough so that common lines based algorithms can succeed. It has not yet been possible to estimate 3D ab-initio models from raw images directly, without class averaging.

Our main motivation in this dissertation is to suggest algorithms for 3D homology modeling that do not require any class averaging, and that can succeed even for small complexes when images are very noisy. In this regime, other existing methods of refinement, class averaging, common lines approaches, fail. The suggested algorithms for 3D homology and ab-initio modeling in this dissertation are based on the theory developed by Zwi Kam [35], and combine experimental design with innovative mathematical techniques. Kam showed in 1980 that the autocorrelation function of the 3D molecule over the rotation group $SO(3)$ can be estimated from 2D projection images whose viewing directions are uniformly distributed over the sphere.

The main requirement for the success of these algorithms is for the number of images in the dataset to be large enough for PCA of the 2D projection images to give non-trivial eigenvalues, that is, a visible gap in the spectrum of the covariance matrix. Covariance estimation in the presence of noise and CTF's is a prerequisite to using Kam's method. In chapter 3, we first derive a new approach for covariance matrix estimation that takes the CTF and noise both into account.

Another requirement for Kam's theory to succeed is for the viewing angles of images to be uniformly distributed over the sphere. We investigate the effects of deviations from this in more detail in chapter 4.

In this dissertation, we focus on 3D ab-initio and homology modeling, and image restoration. In that spirit, we treat datasets as belonging to a single conformation (which is a good approximation when the dataset is largely homogeneous). In cases when the molecule may exhibit heterogeneity, that is the molecule can exist in two or more conformations, this means obtaining its 'average' 3D conformation.

1.4 Image Restoration

Cryo-EM images suffer from degradation due to both electron noise and the effect of the microscope's point spread function, called the Contrast Transfer Function (CTF). The CTF is roughly a radially isotropic decaying sinusoid function in frequency space that is pointwise multiplied with the image in Fourier space. This results in inversion of the contrast where the CTF is negative, and complete loss of information at the 'zero frequencies', which are the frequencies at which the CTF is zero. CTF correction is a challenging problem due to this non invertible nature of the CTF (see Fig. 1.3). Phase flipping [58] is a popular albeit sub-optimal technique that corrects only for the phases but not for the amplitudes of the Fourier coefficients. In phase flipping, the sign of Fourier coefficients is inverted at frequencies for which the CTF is negative. There exist other approaches such as classical Wiener filtering. Wiener filtering can be considered as careful division or inversion of the CTF operator so as to limit the accompanying noise amplification. Wiener filtering gives a linear estimate of the image that is optimal in terms of the mean squared error. In some approaches, CTF correction is handled much later in the cryo-EM pipeline at the 3D reconstruction stage [66, 103, 96, 82]. Cryo-EM micrographs are acquired at different microscope settings such as the defocus, so different images are affected by different CTF's. An optimal approach for CTF correction should correct for both the Fourier phases and amplitudes. This would require combining information from all images that are

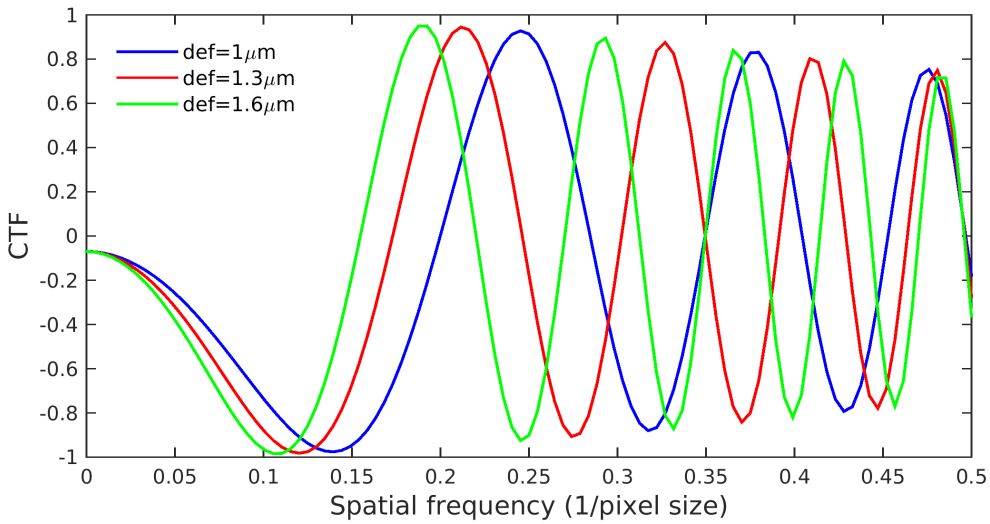


Figure 1.3: CTF's for different values of the defocus. CTF parameters used are: the amplitude contrast $\alpha = 0.07$, the electron wavelength $\lambda = 2.51pm$, the spherical aberration constant $Cs = 2.0$, the B-factor $B = 10$, the defocus= $1\mu m, 1.3\mu m$, and $1.6\mu m$, and the pixel size is $2.82\mathring{A}$.

affected by different CTF's, to recover information lost at the zero crossings.

Image restoration in cryo-EM requires both denoising and CTF correction. Image restoration of the noisy, 2D projection images is required to aid the process of 3D reconstruction itself. Typically, 3D reconstruction is performed starting from class averages which enjoy a higher SNR than raw images. There are several methods to decrease the noise level using low pass filters, wavelet transforms, non-local mean filters, bilateral filters, etc. [105, 33, 92, 108]. The main challenge in denoising is to retain important features in the image while reducing the noise.

1.5 Contributions

In this thesis, we introduce new algorithms for 3D homology and ab-initio modeling in cryo-EM, that leverage Kam's theory and generalize the molecular replacement method in X-ray crystallography. We foresee these new algorithms to enable reconstruction of small complexes for which existing methods of refinement, class averaging and common-lines approaches fail because the images are very noisy. Our algorithms combine mathematical innovation with experimental design, and are less restrictive in terms of the SNR required for other approaches to succeed.

In Chapter 1, we introduce two new algorithms for homology and ab-initio modeling based on Kam's theory. The autocorrelation function determines the expansion coefficients of the 3D molecule in spherical harmonics up to an orthogonal matrix of size $(2l+1)\times(2l+1)$ for each $l = 0, 1, 2, \cdots$. In this chapter we show how techniques for solving the phase retrieval problem in X-ray crystallography can be modified for the cryo-EM setup for retrieving the missing orthogonal matrices. Specifically, we present two new approaches that we term *Orthogonal Extension* and *Orthogonal Replacement*, in which the main algorithmic components are the singular value decomposition and semidefinite programming. We demonstrate the utility of these approaches through numerical experiments on simulated data.

Our homology and ab-initio modeling approaches require an estimator for the covariance matrix of the underlying clean images. The covariance estimator can be leveraged for image restoration. The problem of image restoration in cryo-EM entails correcting for the effects of the Contrast Transfer Function (CTF) and noise. Popular methods for image restoration include 'phase flipping', which corrects only for the Fourier phases but not amplitudes, and Wiener filtering, which requires the spectral signal to noise ratio. In chapter 3, we propose a new image restoration method which we call 'Covariance Wiener Filtering' (CWF). In CWF, the covariance matrix of the projection images is used within the classical Wiener filtering framework for

solving the image restoration deconvolution problem. Our estimation procedure for the covariance matrix is new and successfully corrects for the CTF. We demonstrate the efficacy of CWF by applying it to restore both simulated and experimental cryo-EM images. Results with experimental datasets demonstrate that CWF provides a good way to evaluate the particle images and to see what the dataset contains even without 2D classification and averaging.

One of the main challenges in cryo-EM is the typically low signal to noise ratio (SNR) of the acquired images. 2D classification of images, followed by class averaging, improves the SNR of the resulting averages, and is used for selecting particles from micrographs and for inspecting the particle images. In chapter 4, we introduce a new affinity measure, akin to the Mahalanobis distance, to compare cryo-EM images belonging to different defocus groups. The new similarity measure is employed to detect similar images, thereby leading to an improved algorithm for class averaging. We evaluate the performance of the proposed class averaging procedure on synthetic datasets, obtaining state of the art classification.

Finally, in chapter 5, we utilize the covariance matrix estimator from CWF and integrate it into the Orthogonal Retrieval procedure. The missing phase problem in X-ray crystallography is commonly solved using the technique of molecular replacement [74, 73, 78], which borrows phases from a previously solved homologous structure, and appends them to the measured Fourier magnitudes of the diffraction patterns of the unknown structure. More recently, molecular replacement has been proposed for solving the missing orthogonal matrices problem arising in Kam's autocorrelation analysis [34, 35] for single particle reconstruction using X-ray free electron lasers [75, 32, 93] and cryo-EM [9]. In classical molecular replacement, it is common to estimate the magnitudes of the unknown structure as twice the measured magnitudes minus the magnitudes of the homologous structure, a procedure known as 'twicing' [97]. Mathematically, this is equivalent to finding an unbiased estimator for

a complex-valued scalar [54]. We generalize this scheme for the case of estimating real or complex valued matrices arising in single particle autocorrelation analysis. We name this approach "Anisotropic Twicing" because unlike the scalar case, the unbiased estimator is not obtained by a simple magnitude isotropic correction. We compare the performance of the least squares, twicing and anisotropic twicing estimators on synthetic and experimental datasets. We demonstrate 3D homology modeling in cryo-EM directly from experimental data without iterative refinement or class averaging, for the first time.

Chapter 2

Orthogonal Matrix Retrieval in Cryo-EM

2.1 Introduction

Single Particle Reconstruction (SPR) from cryo-EM is an increasingly popular technique in structural biology for determining 3D structures of macromolecular complexes that resist crystallization [23, 44, 5]. In the basic setup of SPR, the data collected are 2D projection images of ideally assumed identical, but randomly oriented, copies of a macromolecule. In cryo-EM, the sample of molecules is rapidly frozen in a thin layer of vitreous ice, and maintained at liquid nitrogen temperature throughout the imaging process [106]. The electron microscope provides a top view of the molecules in the form of a large image called a micrograph. The projections of the individual particles can be picked out from the micrograph, resulting in a set of projection images. Datasets typically range from 10^4 to 10^5 projection images whose size is roughly 100×100 pixels.

Mathematically, ignoring the effects of the microscope's contrast transfer function and noise, a 2D projection image $I : \mathbb{R}^2 \to \mathbb{R}$ corresponding to rotation R is given by

the integral of the Coulomb potential $\varphi : \mathbb{R}^3 \to \mathbb{R}$ that the molecule induces

$$I(x, y) = \int_{-\infty}^{\infty} \varphi(R^T r)\, dz, \tag{2.1}$$

where $r = (x, y, z)^T$. The 3D reconstruction problem in cryo-EM is a non-linear inverse problem in which φ needs to be estimated from multiple noisy discretized projection images of the form (2.1) for which the rotations are unknown.

Radiation damage limits the maximum allowed electron dose. As a result, the acquired 2D projection images are extremely noisy with poor signal-to-noise ratio (SNR). Estimating φ and the unknown rotations at very low SNR is a major challenge.

The 3D reconstruction problem is typically solved by guessing an initial structure and then performing an iterative refinement procedure, where iterations alternate between estimating the rotations given a structure and estimating the structure given rotations [23, 79, 64]. When the particles are too small and images too noisy, the final result of the refinement process depends heavily on the choice of the initial model, which makes it crucial to have a good initial model. If the molecule is known to have a preferred orientation, then it is possible to find an *ab-initio* 3D structure using the random conical tilt method [50, 51]. There are two known approaches to ab initio estimation that do not involve tilting: the method of moments [25, 76], and common-lines based methods [99, 29, 88].

Using common-lines based approaches, [115] was able to obtain three-dimensional ab-initio reconstructions from real microscope images of large complexes that had undergone only rudimentary averaging. However, researchers have so far been unsuccessful in obtaining meaningful 3D ab-initio models directly from raw images that have not been averaged, especially for small complexes.

We present here two new approaches for ab-initio modelling that are based on Kam's theory [35] and that can be regarded as a generalization of the molecular

replacement method from X-ray crystallography to cryo-EM. The only requirement for our methods to succeed is that the number of collected images is large enough for accurate estimation of the covariance matrix of the 2D projection images.

2.2 Kam's theory and the Orthogonal matrix retrieval problem

Kam showed [35] using the Fourier projection slice theorem (see, e.g., [59, p. 11]) that if the viewing directions of the projection images are uniformly distributed over the sphere, then the autocorrelation function of the 3D volume with itself over the rotation group SO(3) can be directly computed from the covariance matrix of the 2D images. Let $\hat{\varphi} : \mathbb{R}^3 \to \mathbb{C}$ be the 3D Fourier transform of φ and consider its expansion in spherical coordinates

$$\hat{\varphi}(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(k) Y_l^m(\theta, \varphi) \tag{2.2}$$

where k is the radial frequency and Y_l^m are the real spherical harmonics. Kam showed that

$$C_l(k_1, k_2) = \sum_{m=-l}^{l} A_{lm}(k_1) \overline{A_{lm}(k_2)} \tag{2.3}$$

can be estimated from the covariance matrix of the 2D projection images. For images sampled on a Cartesian grid, each matrix C_l is of size $K \times K$, where K is the maximum frequency (dictated by the experimental setting). In matrix notation, eq.(2.3) can be rewritten as

$$C_l = A_l A_l^*, \tag{2.4}$$

where A_l is a matrix size $K \times (2l + 1)$ whose m'th column is A_{lm}. The factorization (2.4) of C_l, also known as the Cholesky decomposition, is not unique: If A_l satisfies

(5.3), then $A_l U$ also satisfies (5.3) for any $(2l+1) \times (2l+1)$ unitary matrix U (i.e., $UU^* = U^*U = I$).

Since φ, the electric potential induced by the molecule, is real-valued, its Fourier transform $\hat{\varphi}$ satisfies $\hat{\varphi}(r) = \overline{\hat{\varphi}(-r)}$, or equivalently, $\hat{\varphi}(k, \theta, \varphi) = \overline{\hat{\varphi}(k, \pi - \theta, \varphi + \pi)}$. Together with properties of the real spherical harmonics, it follows that $A_{lm}(k)$ (and therefore A_l) is real for even l and purely imaginary for odd l. Then A_l is unique up to a $(2l+1) \times (2l+1)$ orthogonal matrix $O_l \in \mathrm{O}(2l+1)$, where

$$\mathrm{O}(d) = \{O \in \mathbb{R}^{d \times d} \,:\, OO^T = O^T O = I\}. \tag{2.5}$$

Originally, $2l+1$ functions of the radial frequency are required for each l in order to completely characterize φ. With the additional knowledge of C_l the parameter space is reduced to $\mathrm{O}(2l+1)$. We refer to the problem of recovering the missing orthogonal matrices $O_0, O_1, O_2, \ldots$ as the *orthogonal matrix retrieval problem in cryo-EM*.

2.2.1 Analogy with X-ray crystallography

The orthogonal matrix retrieval problem is akin to the phase retrieval problem in X-ray crystallography. In crystallography, the measured diffraction patterns contain information about the modulus of the 3D Fourier transform of the structure but the phase information is missing and needs to be obtained by other means. Notice that in crystallography, the particle's orientations are known but the phases of the Fourier coefficient are missing, while in electron microscopy, the projection images contain phase information but the orientations of the particles are missing. Kam's theory converts the cryo-EM problem to one akin to the phase retrieval problem in crystallography. From a mathematical standpoint, the phase retrieval problem in crystallography is perhaps more challenging than the orthogonal matrix retrieval problem in cryo-EM, because in crystallography each Fourier coefficient is missing its

phase, while in cryo-EM only a single orthogonal matrix is missing per several radial components.

2.3 Orthogonal Extension (OE)

A classical solution to the missing phase problem in crystallography is molecular replacement, which relies upon the existence of a previously solved structure which is similar to the unknown structure from which the diffraction data is obtained. The structure is then estimated using the Fourier magnitudes from the diffraction data with the phases from the homologous structure. We mimic this approach in cryo-EM, by grafting the orthogonal matrices of the already resolved similar structure onto the unknown structure.

Let φ be the unknown structure, and suppose ψ is a known homologous structure, whose 3D Fourier transform $\hat{\psi}$ has the following expansion in spherical harmonics

$$\hat{\psi}(k,\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} B_{lm}(k) Y_l^m(\theta,\varphi) \tag{2.6}$$

We can obtain the auto-correlation matrices C_l from the cryo-EM images of the unknown structure φ using Kam's method. Let F_l be any matrix satisfying $C_l = F_l F_l^*$, determined from the Cholesky decomposition of C_l. Then

$$A_l = F_l O_l \tag{2.7}$$

where $O_l \in \mathrm{O}(2l+1)$. Requiring $A_l \approx B_l$, in *orthogonal extension* we determine O_l as the solution to the least squares problem

$$O_l = \underset{O \in \mathrm{O}(2l+1)}{\arg\min} \|F_l O - B_l\|_F^2, \tag{2.8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Although the orthogonal group is non-convex, there is a closed form solution to (2.8) (see, e.g., [38]) given by

$$O_l = V_l U_l^T, \tag{2.9}$$

where

$$B_l^* F_l = U_l \Sigma_l V_l^T \tag{2.10}$$

is the singular value decomposition (SVD) of $B_l^* F_l$. Thus, we estimate A_l by

$$A_l = F_l V_l U_l^T. \tag{2.11}$$

In analogy with crystallography, the phase information ($V_l U_l^T$) from the resolved homologous structure appends the experimentally measured intensity information (F_l). We note that other magnitude correction schemes have been used in crystallography. For example, setting the magnitude to be twice the magnitude from the desired structure minus the magnitude from the known structure, has the desired effect of properly weighting the difference between the two structures, but also the undesired effect of doubling the noise level. The cryo-EM analog in this case would be estimating A_l by

$$A_l = 2F_l V_l U_l^T - B_l. \tag{2.12}$$

2.4 Orthogonal Replacement (OR)

We move on to describe *Orthogonal Replacement*, our approach for resolving structures for which there does not exist a homologous structure. Suppose $\varphi^{(1)}$ and $\varphi^{(2)}$ are two unknown structures for which we have cryo-EM images. We assume that their difference $\Delta\varphi = \varphi^{(2)} - \varphi^{(1)}$ is known. This can happen, for example, when an antibody fragment of a known structure binds to a protein. Another example is when

a complex is obtained by docking another known complex to an α-helix or a β-sheet whose structure is known to high resolution. We have two sets of cryo-EM images, one from the protein alone, $\varphi^{(1)}$ and another from the protein plus the antibody, $\varphi^{(2)}$. Let $C_l^{(i)}$ be the matrices computed from the sample covariance matrices of the 2D projection images of $\varphi^{(i)}, (i = 1, 2)$. Let $F_l^{(i)}$ be any matrix satisfying $C_l^{(i)} = F_l^{(i)} F_l^{(i)*}$. We have $A_l^{(i)} = F_l^{(i)} O_l^{(i)}$, where $O_l^{(i)} \in \mathrm{O}(2l+1)$. The matrices $O_l^{(i)}$ need to be determined for $i = 1, 2$ and $l = 0, 1, 2, \cdots$. The difference $A_l^{(2)} - A_l^{(1)}$ is known from the 3D Fourier transform of the binding structure $\Delta\varphi$. We have

$$A_l^{(2)} - A_l^{(1)} = F_l^{(2)} O_l^{(2)} - F_l^{(1)} O_l^{(1)} \tag{2.13}$$

2.4.1 Relaxation to a Semidefinite Program

Viewing (2.13) as a system of linear equations, we estimate $O_l^{(1)}$ and $O_l^{(2)}$ using least squares as

$$\min_{O_l^{(1)}, O_l^{(2)} \in \mathrm{O}(2l+1)} \left\| A_l^{(2)} - A_l^{(1)} - F_l^{(2)} O_l^{(2)} + F_l^{(1)} O_l^{(1)} \right\|_F^2. \tag{2.14}$$

The number of free parameters associated with an orthogonal matrix $\in \mathrm{O}(2l+1)$ is $\binom{2l+1}{2} = l(2l+1)$. But the least squares formulation in (2.14) does not constrain $O_l^{(1)}$ and $O_l^{(2)}$ to $\in \mathrm{O}(2l+1)$. It instead allows them to be any square real valued matrix of size $2l + 1 \times 2l + 1$. The effective total number of variables for the least squares problem in (2.14) is thus $2(2l+1)^2$. The total number of linear equations in (2.13) is $(2l+1)K$. As we need the number of equations to exceed the number of variables, using least sqaures we can resolve a truncated spherical harmonic expansion only for angular frequencies l satisfying $(2l+1)K \geq 2(2l+1)^2$ or equivalently

$$l \leq \frac{K}{4} - \frac{1}{2}. \tag{2.15}$$

This poses a natural resolution limit on structures that can be resolved using the least

squares method.

Ideally, we should restrict the least squares solution to the orthogonal group since in this manner we can obtain a truncated spherical harmonic expansion for angular frequencies satisfying $(2l+1)K \geq 2l(2l+1)$, that is,

$$l \leq \frac{K}{2} \tag{2.16}$$

We find on comparing (2.15) and (2.16) that such a method would provide structures with higher resolution than possible with the least squares method. The main mathematical obstacle is that (2.14) is a non-convex optimization problem with no closed form solution. One possible approach is an alternating least squares procedure, which is an iterative procedure that alternates between updating $O_l^{(1)}$ and $O_l^{(2)}$. Each iteration reduces the cost function so the iterates converge to a local minimum, though not necessarily the global minimum. Numerical simulations show that the alternating least squares method typically does not converge to the global minimum unless started with a good initialization.

We find $O_l^{(1)}$ and $O_l^{(2)}$ using convex relaxation in the form of semidefinite programming (SDP). We first homogenize (2.13) by introducing a slack unitary variable $O_l^{(3)}$ and consider the augmented linear system

$$(A_l^{(2)} - A_l^{(1)})O_l^{(3)} = F_l^{(2)}O_l^{(2)} - F_l^{(1)}O_l^{(1)} \tag{2.17}$$

If the triplet $\{O_l^{(1)}, O_l^{(2)}, O_l^{(3)}\}$ is a solution to (2.17), then the pair $\{O_l^{(1)}O_l^{(3)^T}, O_l^{(2)}O_l^{(3)^T}\}$ is a solution to the original linear system (2.13). The corresponding least squares problem

$$\min_{\substack{O_l^{(i)} \in \mathrm{O}(2l+1) \\ i=1,2,3}} \left\| (A_l^{(2)} - A_l^{(1)})O_l^{(3)} - F_l^{(2)}O_l^{(2)} + F_l^{(1)}O_l^{(1)} \right\|_F^2 \tag{2.18}$$

is still non-convex. But it can be relaxed to an SDP. Let $Q \in \mathbb{R}^{3(2l+1)\times 3(2l+1)}$ be a

symmetric matrix, which can be expressed as a 3×3 block matrix with block size $2l+1$, and the ij'th block is given by

$$Q_{ij} = O_l^{(i)} O_l^{(j)^T}, \quad i,j = 1,2,3 \tag{2.19}$$

It follows that Q is positive semidefinite (denoted $Q \succeq 0$). Moreover, the three diagonal blocks of Q are $Q_{ii} = I$ $(i = 1,2,3)$ and $\text{rank}(Q) = 2l+1$. The cost function in (2.18) is quadratic in $O_l^{(i)} (i = 1,2,3)$, so it is linear in Q. The problem can be equivalently rewritten as

$$\min_{Q} \text{Tr}(WQ) \tag{2.20}$$

over $Q \in \mathbb{R}^{3(2l+1)\times 3(2l+1)}$, subject to $Q_{ii} = I$, $\text{rank}(Q) = 2l+1$ and $Q \succeq 0$, where the matrix W can be written in terms of $A_l^{(2)} - A_l^{(1)}$, $F_l^{(1)}$ and $F_l^{(2)}$. Here, we have only one non-convex constraint – the rank constraint. Upon dropping the rank constraint we arrive at an SDP that can be solved efficiently in polynomial time in l. We extract the orthogonal matrices $O_l^{(i)}$ from the decomposition (2.19) of Q. If the solution matrix Q has rank greater than $2l+1$ (which is possible since we dropped the rank constraint), then we employ the rounding procedure of [7] to find the closest orthogonal matrix via singular value decomposition.

2.4.2 Exact Recovery and Resolution Limit

We have the following theoretical guarantee on recovery of $O_l^{(1)}$ and $O_l^{(2)}$ using the SDP relaxation in the noiseless case:

Theorem 2.4.1. *Assume that $A_l^{(1)}$ and $A_l^{(2)} \in \mathbb{R}^{K\times(2l+1)}$ are elementwise sampled from i.i.d. Gaussian $N(0,1)$, and $K > 2l+1$, then the SDP method recovers $O_l^{(1)}$ and $O_l^{(2)}$ almost surely.*

The proof of Theorem 2.4.1 is detailed in [111]. Theorem 2.4.1 shows that the

SDP method almost achieves the theoretical information limit, since by counting the degrees of freedom in (2.13) it is impossible to recover $O_l^{(1)}$ and $O_l^{(2)}$ if $K < 2l$. Indeed, the number of free parameters associated with an orthogonal matrix in $\mathrm{O}(2l+1)$ is $l(2l+1)$, while the number of equations in (2.13) is $K(2l+1)$. This introduces a natural resolution limit on structures that can be resolved. Only angular frequencies for which $l \leq \frac{K}{2}$ can be determined using OR.

2.5 Numerical experiments

We present the results of numerical experiments on simulated images (109×109 pixels) of the Kv1.2 potassium channel complex (Fig. 2.1 A and B) with clean and noisy projection images. The experiments were performed in MATLAB in UNIX environment on an Intel (R) Xeon(R) X7542 with 2 CPUs, having 6 cores each, running at 2.67 GHz, and with 256 GB RAM in total. To solve the SDP we used the MATLAB package CVX [26], and to compute the covariance matrix of the 2D images we used the steerable PCA procedure [114].

Kv1.2 is a dumbbell-shaped particle consisting of two subunits - a small β_4 subunit and a larger α_4 subunit, connected by a central connector. We performed experiments using OE and OR, assuming one of the subunits (e.g., α_4) is known, while the other is unknown. In the case of OR, we additionally used projection images of the unknown subunit.

2.5.1 Clean and Noisy Projections

We reconstruct the structure from both clean and noisy projection images. The reconstruction of Kv1.2 obtained from clean images using OE and OR is shown in Fig. 2.1 C through F. We used the true C_l matrices for the known subunit, and a maximum l of 30. We tested OR to reconstruct Kv1.2 from noisy projections at

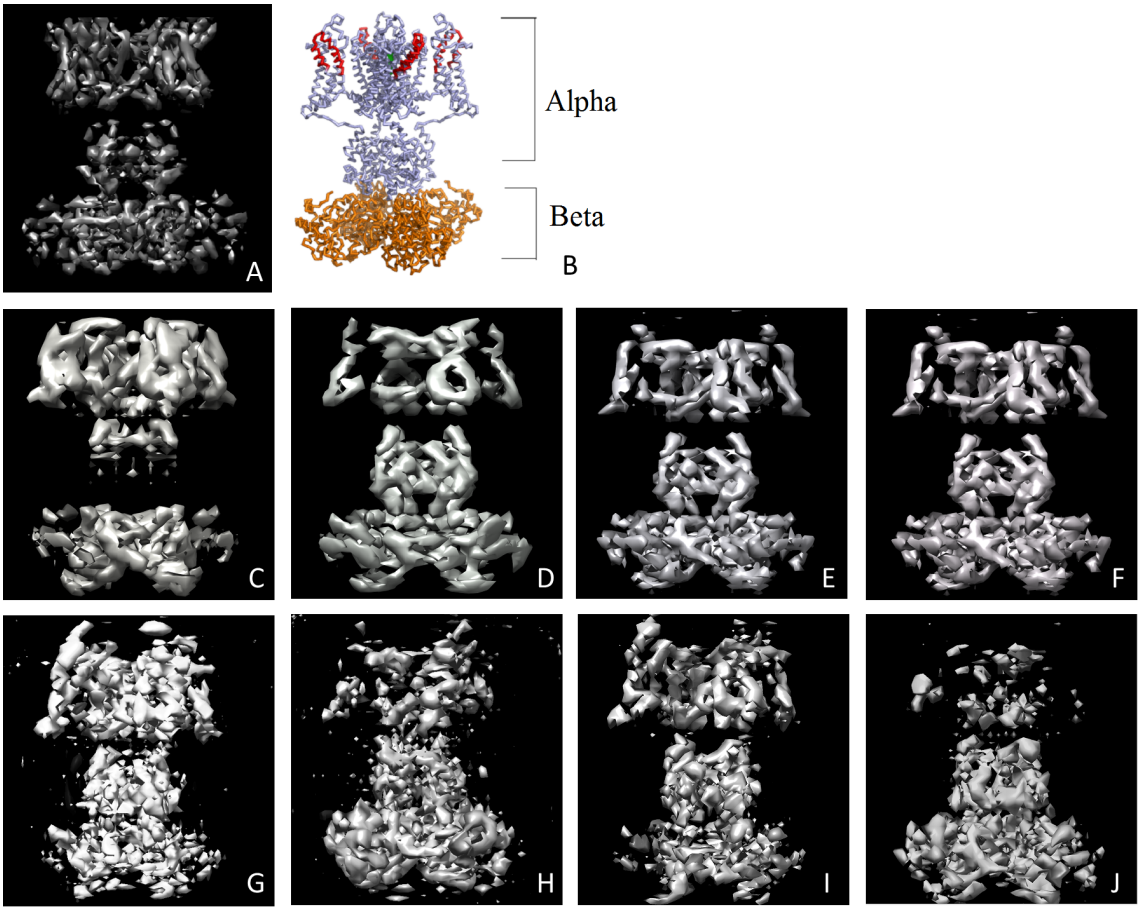


Figure 2.1: Kv1.2 potassium channel: A) Volume visualization in UCSF Chimera [69]. B) Image from Protein Data Bank Japan (PDBj). C through F show reconstructions from clean images - C) OE with α_4 known, D) OE with β_4 known, E) OR with α_4 known, and F) OE with β_4 known. G through J show reconstructions from noisy images using OR - G) SNR=0.7 with α_4 known, H) SNR=0.7 with β_4 known, I) SNR=0.35 with α_4 known, and J) SNR=0.35 with β_4 known.

various values of SNR. A sample projection image at different values of SNR is shown in Fig. 2.2. The C_l matrices were estimated from the noisy projection images. In Fig. 2.1 G through J we show the reconstructions obtained from 10000 projections using OR at SNR=0.7, and from 40000 projections using OR at SNR=0.35. In our simulations with 10000 images, it takes 416 seconds to perform steerable PCA, 194 seconds to calculate the C_l matrices using the maximum l as 30, and the time to solve the SDP as a function of l ranges from 5 seconds for $l = 5$ to 194 seconds for $l = 30$.

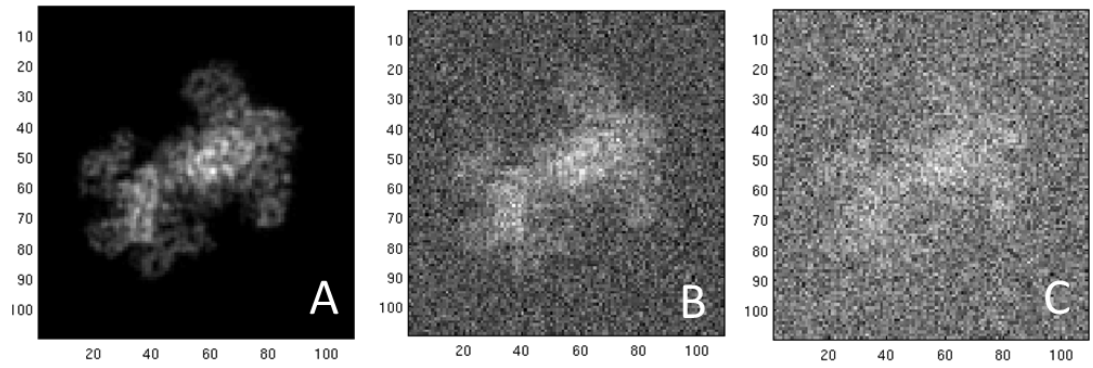


Figure 2.2: Projection images at different values of SNR: A) Clean image, B) SNR=0.7, and C) SNR=0.35.

2.5.2 Comparison between OE and OR

We quantify the 'goodness' of the reconstruction using the Fourier Cross Resolution (FCR) [65]. In Fig. 2.3 we show the FCR curves for the reconstruction from the β_4 complex using OE and OR. The additional information in OR, from the projection images of α_4, results in a better reconstruction, as seen from the FCR curve. The Kv1.2 complex has C4 symmetry, which reduces the rank of the C_l matrices. Our experiment thus benefits from the reduced size of the orthogonal matrices to be recovered.

2.6 Summary

We presented two new approaches based on Kam's theory for *ab-initio* modeling of macromolecules for SPR from cryo-EM. Ab-initio modelling of small complexes is a challenging problem in cryo-EM because it is difficult to detect common lines between noisy projection images at low SNR. Our methods only require reliable estimation of the covariance matrix of the projection images which can be met even at low SNR if the number of images is sufficiently large. We need to first estimate the covariance matrix of 2D projection images prior to the effect of the CTF. One possible approach is to apply CTF correction such as phase flipping to the images and use the

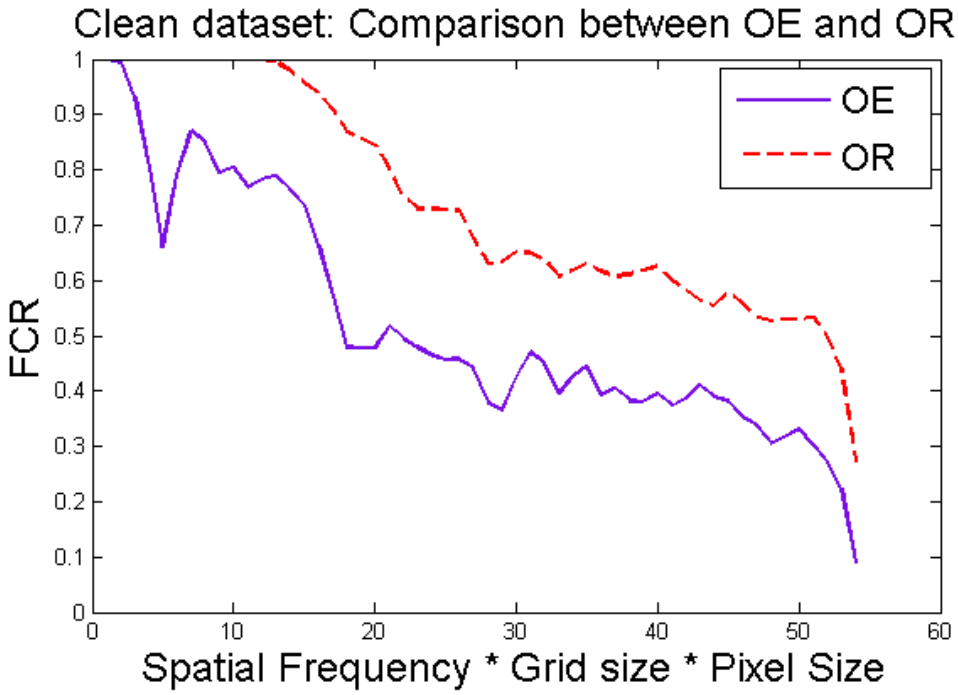


Figure 2.3: FCR curve for reconstruction from β_4 (clean images)

CTF-corrected images to compute the covariance matrix. However, phase flipping correction is sub-optimal since it does not correct the Fourier amplitudes. Instead, we design an optimal estimator of the 2D covariance matrix and combine it with the steerable PCA [114, 113] procedure that performs better than multivariate statistical analysis. This would lead to better 2D image restoration and class averaging procedures. In the next chapter, we estimate the covariance matrix in the presence of the CTF and realistic noise levels, and to apply our methods to experimental datasets.

Chapter 3

Denoising and Covariance Estimation of Single Particle Cryo-EM Images

3.1 Introduction

Single particle reconstruction (SPR) using cryo-electron microscopy (cryo-EM) is a rapidly advancing technique for determining the structure of biological macromolecules at near-atomic resolution directly in their native state, without any need for crystallization [5, 58, 60, 85, 43]. In SPR, 3D reconstructions are estimated by combining multiple noisy 2D tomographic projections of macromolecules in different unknown orientations.

The acquired data consists of multiple micrographs from which particle images are extracted in the first step of the computational pipeline. Next, the images are grouped together by similarity in the 2D classification and averaging step [115, 62]. Class averages can be used to inspect the underlying particles, and to estimate viewing angles and form a low resolution ab-initio 3D model. Subsequently, this 3D model is

refined to high resolution, and 3D classification might be performed as well.

In this chapter, we propose an image restoration method that provides a way for visualizing the particle images without performing any 2D classification. While noise reduction is achieved in 2D classification by averaging together different particle images, our method operates on each image separately, and performs contrast transfer function (CTF) correction and denoising in a single step.

Existing image restoration techniques (for denoising and CTF correction) can be broadly categorized into two kinds of approaches [66]. The first is an approach known as 'phase flipping', which involves flipping the sign of the Fourier coefficients at frequencies for which the CTF is negative. Consequently, phase flipping restores the correct phases of the Fourier coefficients, but ignores the effect of the CTF on the amplitudes. Phase flipping preserves the noise statistics and is easy to implement, leading to its widespread usage in several cryo-EM software packages. However, it is suboptimal because it does not restore the correct Fourier amplitudes of the images. The second commonly used approach is Wiener filter based restoration, to which we refer here as traditional Wiener filtering (TWF). Wiener filtering takes into account both the phases and amplitudes of the Fourier coefficients, unlike phase flipping. However, calculation of the Wiener filter coefficients requires prior estimation of the spectral signal to noise ratio (SSNR) of the signal, which by itself is a challenging problem. It is therefore customary to either treat the SSNR as a precomputed constant as in the software package SPIDER [82], or to apply Wiener filtering only at later stages of the 3D reconstruction pipeline when the noise level is sufficiently low, such as in EMAN2 [96]. It is also possible to use a combination of the two approaches, by first phase flipping the 2D images, and later correct only for the amplitudes in the 3D reconstruction step, as in IMAGIC [103, 102]. Despite its simplicity, there are several drawbacks to TWF. First, it cannot restore information at the zero crossings of the CTF. Second, it requires estimation of the SSNR. Third, it is restrictive to the

Fourier basis which is a fixed basis not adaptive to the image dataset.

We refer to our proposed method as Covariance Wiener Filtering (CWF). CWF consists of first estimating the CTF-corrected covariance matrix of the underlying clean 2D projection images, followed by application of the Wiener filter to denoise the images. Unlike phase flipping, CWF takes into account both the phases and magnitudes of the images. Moreover, unlike TWF that always operates in the data-independent Fourier domain, CWF is performed in the data-dependent basis of principal components (i.e., eigenimages). Crucially, CWF can be applied at preliminary stages of data processing on raw 2D particle images. The resulting denoised images can be used for an early inspection of the dataset, to identify the associated symmetry, and to eliminate 'bad' particle images prior to 2D classification and 3D reconstruction. Additionally, the estimation of the 2D covariance matrix is itself of interest, for example, in Kam's approach for 3D reconstruction [35, 9] (see chapter 2).

The chapter is organized as follows: sections 3.2.1 and 3.2.2 detail the estimation of the covariance matrix for two different noise models, first for the simpler model of white noise, and second for the more realistic model of colored noise. In section 3.2.3 we discuss the steerability property of the covariance matrix [114]. The associated deconvolution problem is solved to obtain denoised images using the estimated covariance matrix in section 3.2.4. Finally in section 3.3, we demonstrate CWF in a number of numerical experiments, with both simulated and experimental datasets. We obtain encouraging results for experimental datasets, in particular, those acquired with the modern direct electron detectors. Image features are clearly observed after CWF denoising. For reproducibility, the MATLAB code for CWF and its dependencies are available in the open source cryo-EM toolbox ASPIRE at `www.spr.math.princeton.edu`.

3.2 Methods

The first step of CWF is estimation of the covariance matrix of the underlying clean images, to which we refer as the population covariance. The second step of CWF is solving a deconvolution problem to recover the underlying clean images using the estimated covariance. In the rest of this section, we describe these steps in detail.

3.2.1 The Model

The image formation model in cryo-EM under the linear, weak phase approximation [22] is given by

$$y_i = a_i * x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{3.1}$$

where n is the number of images, $*$ denotes the convolution operation, y_i is the noisy, CTF filtered i'th image in real space, x_i is the underlying clean projection image in real space, a_i is the point spread function of the microscope that convolves with the clean image in real space, and ϵ_i is additive Gaussian noise that corrupts the image, for each i. Taking the Fourier transform of eqn. 3.1 gives

$$Y_i = A_i X_i + \xi_i, \quad i = 1, 2, \ldots, n \tag{3.2}$$

where Y_i, X_i and ξ_i are now in Fourier space. A_i is a diagonal operator, whose diagonal consists of the Fourier transform of the point spread function, and is also commonly known as the CTF. The CTF modulates the phases and the amplitudes of the Fourier coefficients of the image, and contains numerous zero crossings that correspond to frequencies at which no information is obtained. Any image restoration technique that aims to completely correct for the CTF must therefore correctly restore both the phases and the amplitudes. The zero crossings make CTF correction challenging

since it cannot be trivially inverted. In experiments, different groups of images are acquired at different defocus values, in the hope that information that is lost from one group could be recovered from another group that has different zero crossings. In the experimental datasets used in this paper, the number of images per defocus group typically ranges from 50 to 1000.

In our statistical model, the Fourier transformed clean images $X_1, \ldots, X_n$ (viewed, for mathematical convenience, as vectors in $\mathbb{C}^p$, where p is the number of pixels) are assumed to be independent, identically distributed (i.i.d.) samples from a distribution with mean $\mathbb{E}[\mathbf{X}] = \mu$ and covariance $\mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \Sigma$. Since the clean images are two-dimensional projections of the three-dimensional molecule in different orientations, the distribution of $\mathbf{X}$ in our model is determined by the three-dimensional structure, the distribution of orientations, the varying contrast due to changes in ice thickness, and structural variability, all of course unknown at this stage. The covariance matrix Σ therefore represents the overall image variability due to these determinants. While these model assumptions do not necessarily hold in reality [90, 91], they simplify the analysis and, as will be shown later lead to excellent denoising. Quoting George Box, "All models are wrong but some are useful" [11].

Our denoising scheme requires μ and Σ. Since these quantities are not readily given, we estimate them from the noisy images themselves as follows. For simplicity, we first assume that the noise in our model is additive white Gaussian noise such that $\xi_i \sim \mathcal{N}(0, \sigma^2 I_{p\times p})$ in eqn. 3.2 are i.i.d. The white noise assumption is later replaced by that of the more realistic colored noise. First, notice from eqn. 3.2 it follows that

$$\mathbb{E}[\mathbf{Y}_i] = A_i \mathbb{E}[\mathbf{X}_i], \quad i = 1, 2, \ldots, n. \tag{3.3}$$

So,

$$\begin{aligned}\mathbb{E}[(\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i])(\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i])^T] &= \mathbb{E}[A_i(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^T A_i^T] + \sigma^2 I \\ &= A_i \Sigma A_i^T + \sigma^2 I.\end{aligned} \tag{3.4}$$

Eqn. 3.4 relates the second order statistics of the noisy images with the population covariance Σ of the clean images, based on which we can estimate Σ.

Next, we construct estimators for the mean μ and population covariance Σ using eqn. 3.3 and 3.4. The mean μ of the dataset can be estimated as the solution to a least squares problem

$$\hat{\mu} = \arg\min \mu \sum_{i=1}^{n} ||(Y_i - A_i\mu)||_2^2 + \lambda ||\mu||_2^2 \tag{3.5}$$

where $\lambda \geq 0$ is a regularization parameter. The solution to 3.5 is explicitly

$$\hat{\mu} = (\sum_{i=1}^{n} A_i^T A_i + \lambda I)^{-1} (\sum_{i=1}^{n} A_i^T Y_i). \tag{3.6}$$

The population covariance Σ can be estimated as

$$\begin{aligned}\hat{\Sigma} &= \arg\min \Sigma \sum_{i=1}^{n} ||(Y_i - \mathbb{E}[\mathbf{Y}_i])(Y_i - \mathbb{E}[\mathbf{Y}_i])^T - (A_i \Sigma A_i^T + \sigma^2 I)||_F^2 \\ &= \arg\min \Sigma \sum_{i=1}^{n} ||A_i \Sigma A_i^T + \sigma^2 I - C_i||_F^2\end{aligned} \tag{3.7}$$

where $C_i = (Y_i - A_i\mu)(Y_i - A_i\mu)^T$ and $||.||_F$ is the Frobenius matrix norm. The estimators $\hat{\mu}$ and $\hat{\Sigma}$ can be shown to be consistent in the large sample limit $n \to \infty$, similar to the result in Appendix B of [37].

To ensure that the estimated covariance is positive semidefinite (PSD), we project it onto the space of PSD matrices by computing its spectral decomposition and retaining only the non negative eigenvalues (and their corresponding eigenvectors). To

solve eqn. 3.7, we differentiate the objective function with respect to Σ and set the derivative to zero. This yields

$$\sum_{i=1}^{n} A_i^T A_i \hat{\Sigma} A_i^T A_i = \sum_{i=1}^{n} A_i^T C_i A_i - \sum_{i=1}^{n} \sigma^2 A_i^T A_i \tag{3.8}$$

Eqn. 3.8 defines a system of linear equations for the elements of the matrix $\hat{\Sigma}$. However, direct inversion of this linear system is slow and computationally impractical for large image sizes. Notice that eqn. 3.8 can be written as

$$L(\hat{\Sigma}) = B \tag{3.9}$$

where $L : \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times p}$ is the linear operator acting on $\hat{\Sigma}$ defined by the left hand side of eqn. 3.8, and B is the right hand side. Since applying L only involves matrix multiplications, it can be computed fast, and the conjugate gradient method is employed to efficiently compute $\hat{\Sigma}$ instead of direct inversion, similar to how it is used in [4].

Notice that $L(\hat{\Sigma})$ is a PSD matrix whenever $\hat{\Sigma}$ is PSD (as a sum of PSD matrices), while B may not necessarily be PSD due to finite sample fluctuations (i.e., n is finite). It is therefore natural to project B onto the cone of PSD matrices. This amounts to computing the spectral decomposition of B and setting all negative eigenvalues to 0, which is an instance of eigenvalue thresholding.

We now describe an alternate eigenvalue thresholding procedure, better suited to cases in which the number of images n is not exceedingly large. To that end, we first analyze the matrix B when $X_i = 0$ for all i, i.e., the input images are white noise images containing no signal. Let

$$M = \sum_{i=1}^{n} A_i^T C_i A_i = \sum_{i=1}^{n} A_i^T Y_i Y_i^T A_i. \tag{3.10}$$

Then, $\mathbb{E}[M] = \sigma^2 \sum_{i=1}^{n} A_i^T A_i$ and $B = M - \mathbb{E}[M]$. Let $S = (\mathbb{E}[M])^{1/2}$, i.e. S is PSD and $\mathbb{E}[M] = S^2$. Then multiplying both sides of eqn. 3.9 with S^{-1} we get

$$S^{-1}L(\hat{\Sigma})S^{-1} = S^{-1}(M - \mathbb{E}[M])S^{-1} = S^{-1}MS^{-1} - I. \tag{3.11}$$

$S^{-1}MS^{-1}$ can be viewed as a sample covariance matrix of n vectors in $\mathbb{R}^p$ whose population covariance is the identity matrix. When p is fixed and n goes to infinity, all eigenvalues of $S^{-1}MS^{-1}$ converge to 1. In practice, however, n and p are often comparable. In the limit $p, n \to \infty$ and $p/n \to \gamma$ with $0 < \gamma < \infty$, the limiting spectral density of the eigenvalues converges to the Marčenko Pastur (MP) distribution [57], given by

$$MP(x) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{\gamma x} 1_{[\gamma_-, \gamma_+]}, \quad \gamma_\pm = (1 \pm \sqrt{\gamma})^2 \tag{3.12}$$

for $\gamma \leq 1$. It is therefore expected that $S^{-1}MS^{-1}$ would have eigenvalues (considerably) larger than 1, even in the pure white noise case. These large eigenvalues should not be mistakingly attributed to signal. In the case of images containing signal (plus noise), eigenvalues corresponding to the signal can only be detected if they reside outside of the support of the MP distribution. We use the method of [42] to determine the number of eigenvalues corresponding to the signal. We then apply the operator norm eigenvalue shrinkage procedure (see [18]) to those eigenvalues, while setting all other eigenvalues to 0. We then use the conjugate gradient method [1] to solve eqn. 3.11 for $\hat{\Sigma}$, with the right hand side replaced with its shrinkage version. We observed in numerical simulations (see Fig. 3.3) that this procedure typically outperforms other shrinkage methods in terms of the accuracy of the estimated covariance matrix.

[1] While L in eqn. 9 is PSD, the new effective operator in the LHS of eqn. 11 is not necessarily PSD in general. In order to use conjugate gradient, we solve the system $S^{-1}L(S^{-1}\Sigma_S S^{-1})S^{-1} = S^{-1}MS^{-1} - I$, where $\Sigma_S = S\Sigma S$, in which the operator acting on Σ_S in the LHS is PSD. Σ is then obtained from the estimated Σ_S.

3.2.2 Covariance Estimation with Colored noise

So far, we assumed additive white Gaussian noise in the image formation process. In reality, the noise in experimental images is colored. That is, in the image formation model in eqn. 3.2, ξ_i is additive colored Gaussian noise. We preprocess the images in order to "whiten" the noise. The noise power spectrum can be estimated, for example, using the pixels in the corners of the noisy projection images. To do this, we first estimate using correlograms the 2D autocorrelation of the corner pixels of the images which contain mostly noise and no signal. These corner pixels are used to estimate the 1D autocorrelation, which is then extended to populate the 2D isotropic autocorrelation. We then calculate the Fourier transform of the 2D autocorrelation, which is the 2D power spectrum of noise. The noisy projection images in Fourier space are multiplied element-wise by the inverse of the estimated power spectral density, also called the whitening filter, so that the noise in the resulting images is approximately white. Let W be the "whitening" filter, such that

$$WY_i = WA_iX_i + W\xi_i, \quad i = 1, 2, \ldots, n \tag{3.13}$$

and $W\xi_i \sim \mathcal{N}(0, \sigma^2 I)$.

Eqn. 3.13 is reminiscent of eqn. 3.2. It is tempting to define a new "effective" CTF as WA_i and estimate Σ following the same procedure as in the case of white noise. However, the linear system akin to eqn. 3.8 for this case is ill-conditioned due to the product of W with the CTF, and it takes a large number of iterations for conjugate gradient to converge to the desired solution. Instead, we seek an approach in which the linear system to solve is well conditioned as that in the case of white noise. Since the CTF's $A_i,\ i = 1, 2, \ldots n$ and the whitening filter W are diagonal operators in the Fourier basis, they commute, and eqn. 3.13 becomes

$$WY_i = A_iWX_i + W\xi_i, \quad i = 1, 2, \ldots, n. \tag{3.14}$$

We therefore absorb W into X_i, and estimate the matrix $\Sigma_W = W\Sigma W^T$ (the population covariance of $W\mathbf{X}$) using the same procedure as before. The population covariance Σ is then estimated as

$$\hat{\Sigma} = W^{-1}\hat{\Sigma}_W(W^T)^{-1}. \tag{3.15}$$

3.2.3 Fourier-Bessel Steerable PCA

The population covariance matrix Σ must be invariant under in-plane rotation of the projection images, therefore it is block diagonal in any steerable basis in which the basis elements are outer products of radial functions and angular Fourier modes. Following [114], we choose to represent the images in a Fourier-Bessel basis and it suffices to estimate each diagonal block $\Sigma^{(k)}$, corresponding to the angular frequency k, separately. The Fourier-Bessel basis [114, 113] consists of p_k basis functions (that satisfy the sampling criterion) for each angular frequency k, where p_k decreases with increasing k. The matrix $\Sigma^{(k)}$ is thus of size $p_k \times p_k$.

An important property of the CTF's A_i and the whitening filter W is that they are radially isotropic [2]. Therefore, the CTF's and the whitening filter are also block diagonal in the Fourier Bessel basis. Eqn. 3.8 (and its analog in the case of colored noise) is hence solved separately for each k to estimate $\Sigma^{(k)}$.

3.2.4 Wiener Filtering

The estimated covariance is further used to solve the associated deconvolution problem in eqn. 3.2 using Wiener filtering. The result is a denoised, CTF corrected image for each noisy, CTF affected measurement Y_i for $i = 1, 2, \ldots n$. We estimate X_i in

[2] In the case of astigmatism, where the CTF deviates slightly from radial isotropy, this is a good approximation to obtain low resolution denoised images.

the white noise model using the Wiener filtering procedure as

$$\hat{X}_i = (I - H_i A_i)\hat{\mu} + H_i Y_i \tag{3.16}$$

where $H_i = \hat{\Sigma} A_i^T (A_i \hat{\Sigma} A_i^T + \sigma^2 I)^{-1}$ is the linear Wiener filter [52]. In the case of colored noise,

$$\hat{X}_i = (I - H_i W A_i)\hat{\mu} + H_i Y_i \tag{3.17}$$

with $H_i = \hat{\Sigma} A_i^T W^T (W A_i \hat{\Sigma} A_i^T W^T + \sigma^2 I)^{-1}$. Since the estimated covariance is block-diagonal in the Fourier Bessel basis, the Wiener filtering procedure is applied to the Fourier Bessel coefficients of the noisy images Y_i for each angular frequency k separately. The denoised Fourier Bessel expansion coefficients are used to reconstruct denoised images in Fourier space that are inverse Fourier transformed to acquire images in real space on a Cartesian grid.

3.2.5 Computational Complexity

In practice, instead of each image being affected by a distinct CTF, all images within a given defocus group have the same CTF. So, given D defocus groups with d_i images in group i, one can equivalently minimize the objective function $\sum_{i=1}^{D} d_i || (A_i \Sigma A_i^T + \sigma^2 I) - \sum_{j=1}^{d_i} \frac{1}{d_i} (Y_{i_j} - \mathbb{E}[\mathbf{Y}_{i_j}])(Y_{i_j} - \mathbb{E}[\mathbf{Y}_{i_j}])^T ||_F^2$ in eqn. 3.7 (here A_i denotes the CTF of the i'th defocus group, and i_j index images in that group). As a result, the sums in eqn. 3.8 range from 1 to D instead of from 1 to n, thereby reducing the computational cost of some operations. For images of size $L \times L$, estimating the mean using eqn. 3.6 takes $O(nL^2)$ (since A_i is diagonal in the Fourier basis for each i). Computing the Fourier Bessel expansion coefficients takes $O(nL^3)$, as detailed in [114, 113]. When solving the linear system in eqn. 3.8 to estimate each $\Sigma^{(k)}$ separately, the matrices in eqn. 3.8 are of size $p_k \times p_k$. It is shown in [114] that $\sum_k p_k = O(L^2)$, $\sum_k p_k^2 = O(L^3)$, and $\sum_k p_k^3 = O(L^4)$. While solving eqn. 3.9 using conjugate gradient

for a given angular frequency, computing the action of the linear operation L on $\Sigma^{(k)}$ takes $O(Dp_k^3)$ per iteration, while computing B takes $O(Dp_k^3 + np_k^2)$. Thus, each iteration of conjugate gradient takes $O(D\sum_k p_k^3)$, that is, $O(DL^4)$ and there is also a one time computation of $O(nL^3)$. Wiener filtering the Fourier Bessel coefficients of an image for a given angular frequency k takes $O(p_k^2)$. So the overall complexity for Wiener filtering the coefficients of all images is $O(nL^3)$. In summary, the overall complexity for CWF is $O(TDL^4 + nL^3)$, where T is the number of conjugate gradient iterations.

3.3 Results

In this section, we apply our algorithm to synthetic and experimental datasets to obtain denoised images. All algorithms are implemented in the UNIX environment, on a machine with 60 cores, running at 2.3 GHz, with total RAM of 1.5TB. We perform numerical experiments with (i) a synthetic dataset with additive white and colored Gaussian noise and (ii) four experimental datasets, two of which were acquired with older detectors, and the other two with state-of-the-art direct electron detectors. For all the experimental datasets, the corresponding estimated CTF parameters were provided with the dataset. For all simulations, we use centered projection images. The algorithm does not require centered images. However, having non-centered images would result in an additional 'blurring' effect in the denoised images.

3.3.1 Simulated Noisy Dataset with White Noise

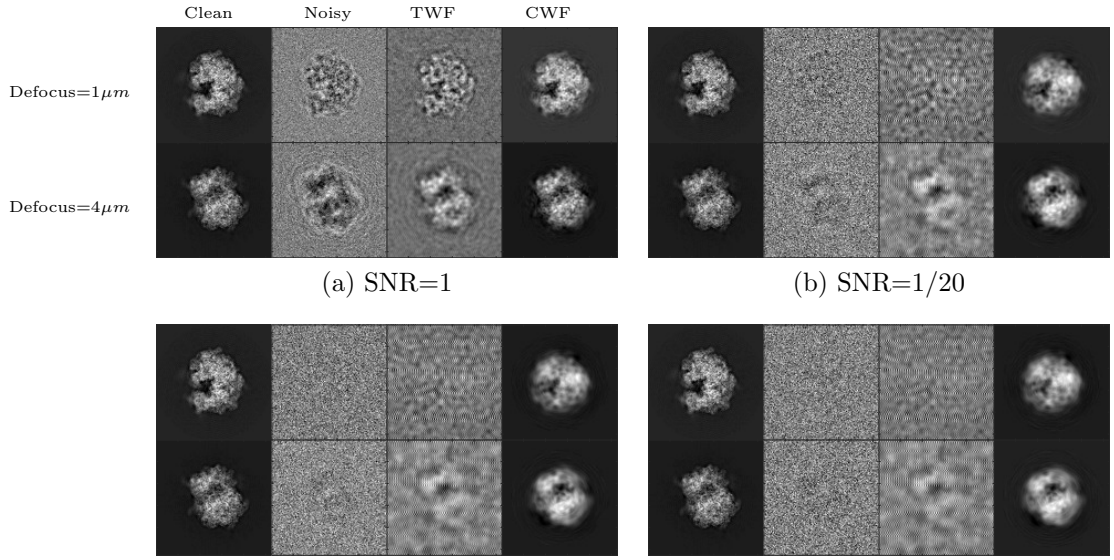
For the first experiment with simulated data, we construct a synthetic dataset by modeling the image formation process in cryo-EM. The synthetic dataset is prepared from the 3D structure of the P. falciparum 80S ribosome bound to E-tRNA, available on the Electron Microscopy Data Bank (EMDB) as EMDB-6454. We first generate

clean 2D projection images starting from a 3D volume, at directions sampled uniformly over the sphere, and then corrupt the generated clean projection images with different CTF's and additive white Gaussian noise. The projection images are divided into 10 defocus groups, with the defocus value ranging from $1\mu m$ to $4\mu m$. The B-factor of the decay envelope was chosen as $10\mathring{A}^2$, the amplitude contrast as 7%, the voltage as 300kV, and the spherical aberration as 2mm. To ensure that the denoising quality of CWF is robust to the mean estimation of the dataset, the regularization parameter λ in the least squares mean estimation in eqn. 3.6 was fixed at 1 for all the experiments described here.

Figure 3.1 shows the results of denoising raw, CTF-affected noisy images with CWF and TWF at various levels of the SNR. We have used the EMAN2 [96] implementation of TWF (note that we perform phase flipping followed by TWF only on the raw images in EMAN2, and not on averages). The SNR used here is defined relative to the CTF affected images that constitute the clean signal, and is calculated as an average value for the entire dataset. Using 20 cores, calculating the Fourier Bessel coefficients took 79 seconds while covariance estimation and Wiener filtering together took 6 seconds in the experiment with SNR= 1/60.

It is seen that TWF works very well at high SNR (≥ 1), but deteriorates at lower SNR's as expected. Note that the denoising results of TWF depend strongly on the defocus value. The location of the zeros in the CTF is such that images corresponding to high defocus values preserve low frequency information, while images corresponding to low defocus values retain more high frequency information. With CWF, there is no such strong dependence on the defocus value, since the covariance matrix is estimated using information from all defocus groups.

Figure 3.2a shows the relative MSE of denoised images as a function of the SNR of the dataset. The MSE (norm of the difference between the denoised image and the original, clean image) shown here corresponds to the same range of SNR's (from 1/60



(a) SNR=1 (b) SNR=1/20

(c) SNR=1/40 (d) SNR=1/60

Figure 3.1: **Synthetic white noise**: A comparison of the denoising results of traditional Wiener filtering (TWF) and CWF for the synthetic dataset prepared from EMDB-6454, the P. falciparum 80S ribosome bound to E-tRNA. The dataset consists of 10000 images of size 105×105, which are divided into 10 defocus groups, with the defocus value ranging from $1\mu m$ to $4\mu m$. The two rows in each subfigure correspond to two clean images belonging to different defocus groups; the first one belongs to the group with the smallest defocus value of $1\mu m$, while the second image belongs to the group with the largest defocus value of $4\mu m$.

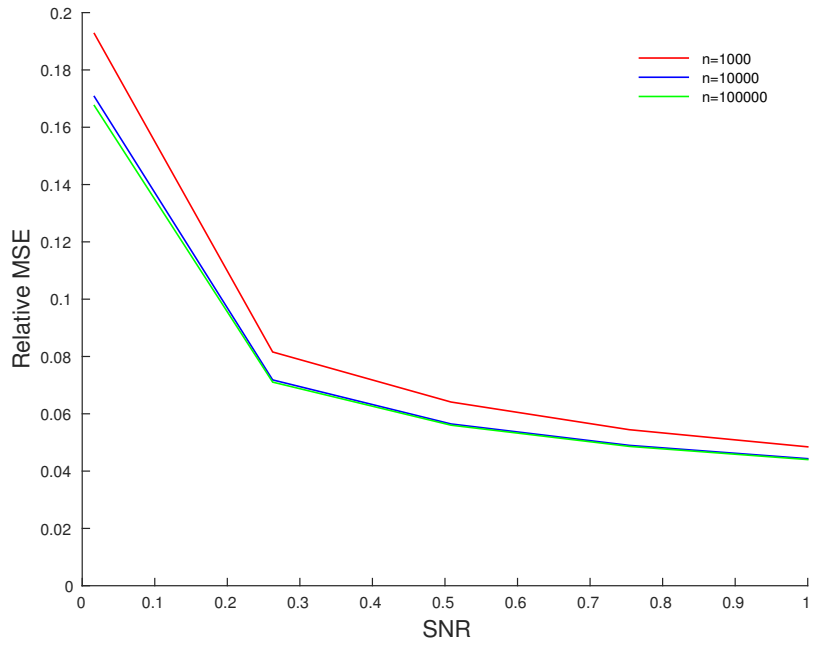
to 1) as in Figure 3.1. Figure 3.2b shows the relative MSE of the denoised images as a function of the number of images used to estimate the covariance in the experiment. The covariance estimation improves as the number of images in the dataset increases, and so the denoising is also expected to improve, as seen from Figure 3.2b.

The importance of the eigenvalue shrinkage procedure is elucidated in Figure 3.3. Here, we compare the error in the estimated covariance with and without eigenvalue shrinkage, for varying number of images used in the experiment. The relative MSE of the estimated covariance $\hat{\Sigma}$ is defined as

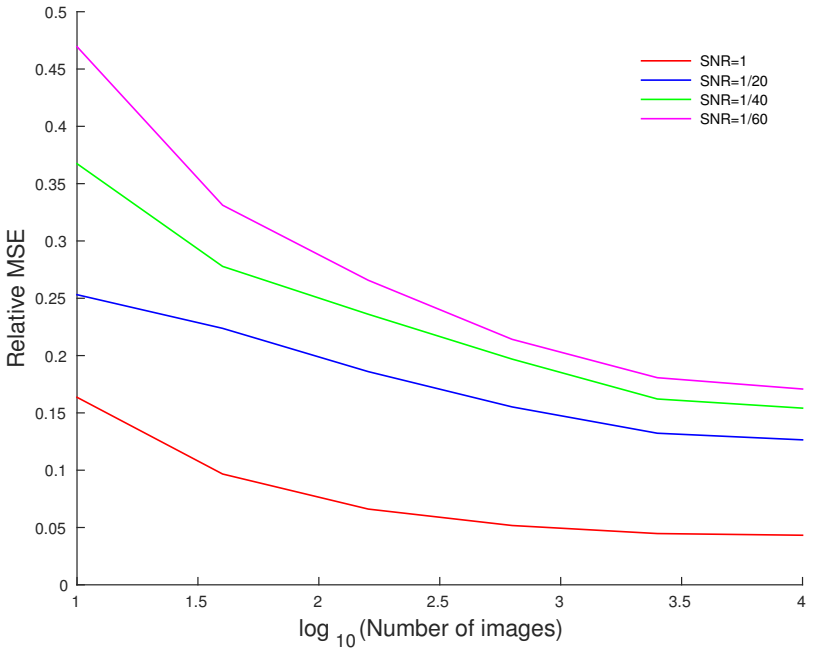
$$MSE_{rel} = \frac{||\Sigma - \hat{\Sigma}||_F^2}{||\Sigma||_F^2} \tag{3.18}$$

3.3.2 Simulated Noisy Dataset with Colored Noise

The noise that corrupts images in cryo-EM is not perfectly white, but often colored. To simulate this, we perform experiments with synthetic data generated from EMDB-6454 as described in 3.3.1, this time adding colored Gaussian noise with the noise response $f(\mathrm{k}) = \frac{1}{\sqrt{(1+\mathrm{k}^2)}}$ (k is the radial frequency) to each clean, CTF-affected projection image. Figure 3.4 shows the denoised images for this case.



(a)



(b)

Figure 3.2: (a) **Relative MSE versus the SNR, for a fixed number of images**: The relative MSE of the denoised images as a function of the SNR, for synthetic data generated using EMDB-6454. The MSE reported here is averaged over all images. n denotes the number of images used in the experiment.(b) **Relative MSE versus the number of images, for a fixed SNR**: The relative MSE of the denoised images as a function of the number of images, for synthetic data generated using EMDB-6454. The MSE reported here is averaged over all images.

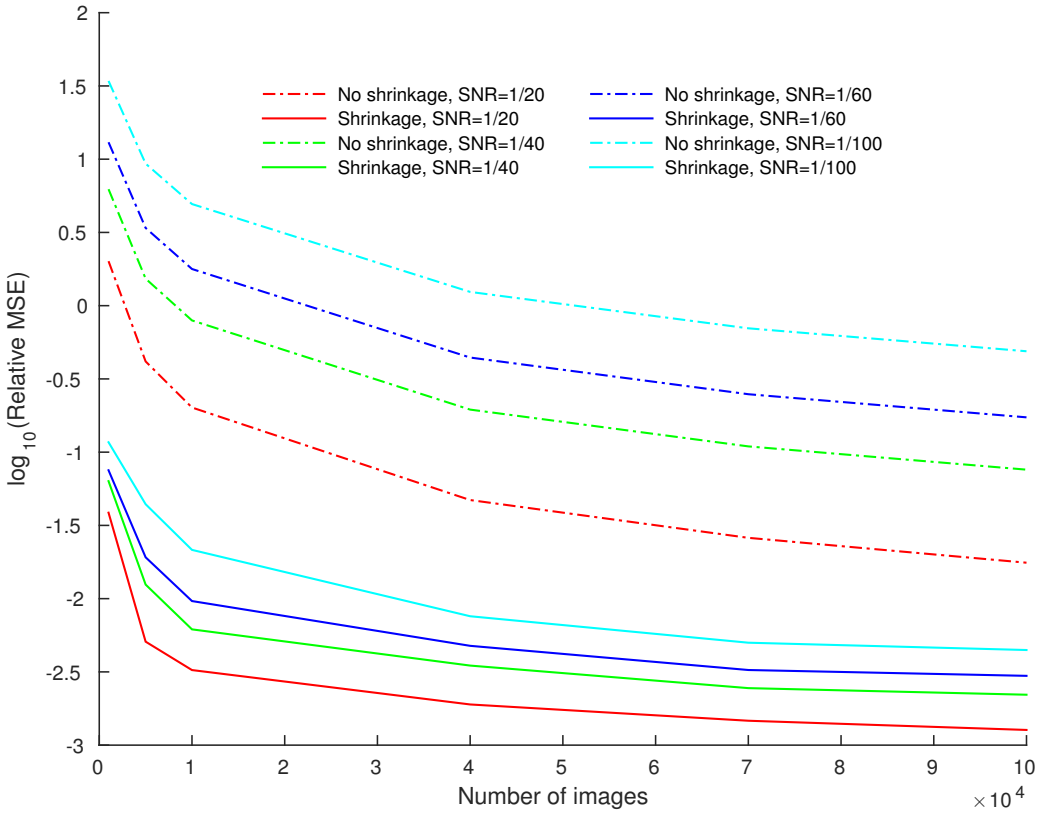
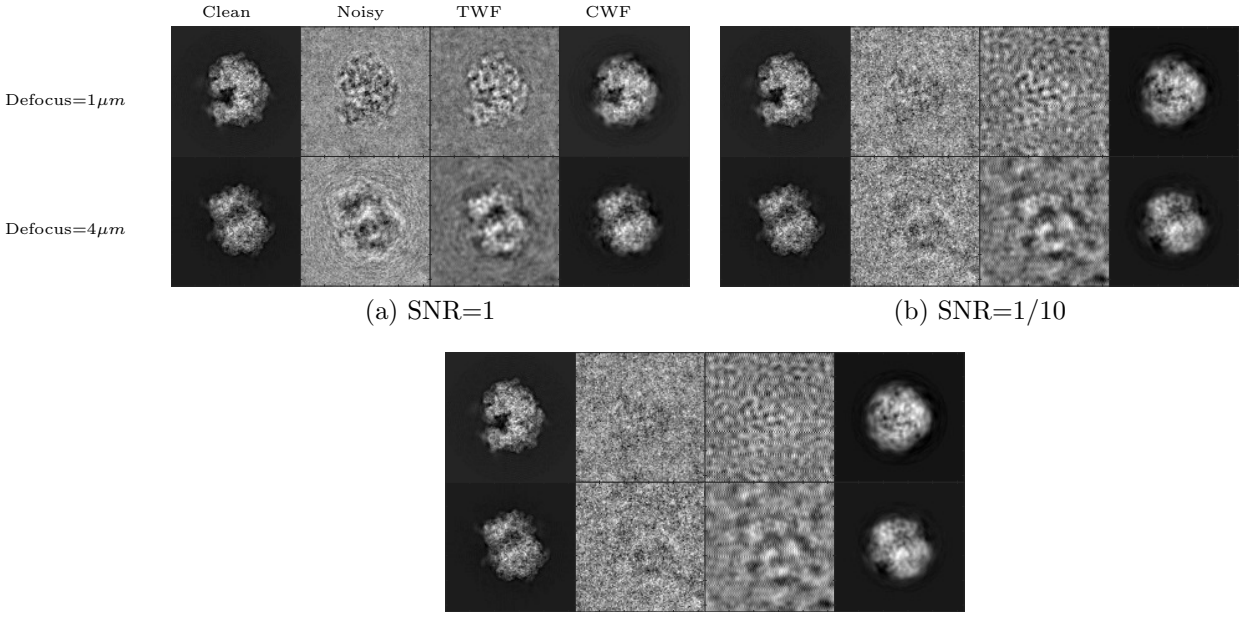


Figure 3.3: **Relative MSE of the estimated covariance versus the number of images**: The relative MSE of the estimated covariance $\hat{\Sigma}$, with and without using eigenvalue shrinkage, as a function of number of images, for synthetic data generated using EMDB-6454.



(a) SNR=1

(b) SNR=1/10

(c) SNR=1/20

Figure 3.4: **Synthetic colored noise**: Denoising results of CWF for the synthetic dataset with additive colored Gaussian noise, prepared from EMDB-6454, the P. falciparum 80S ribosome bound to E-tRNA, as detailed in the caption of Figure 3.1.

3.3.3 Experimental Dataset - TRPV1

We apply CWF to an experimental dataset of the TRPV1 ion channel, taken using a K2 direct electron detector. It is available on the public database Electron Microscope Pilot Image Archive (EMPIAR) as EMPIAR-10005, and the 3D reconstruction is available on EMDB as EMDB-5778, courtesy of Liao et al. [47]. The dataset consists of 35645 motion corrected, picked particle images of size 256×256 pixels with a pixel size of $1.2156\mathring{A}$. Using 20 cores, calculating the Fourier Bessel coefficients took 312 seconds while covariance estimation and Wiener filtering together took 574 seconds. The result is shown in Figure 3.5. CWF retains 384 eigenvalues of Σ.

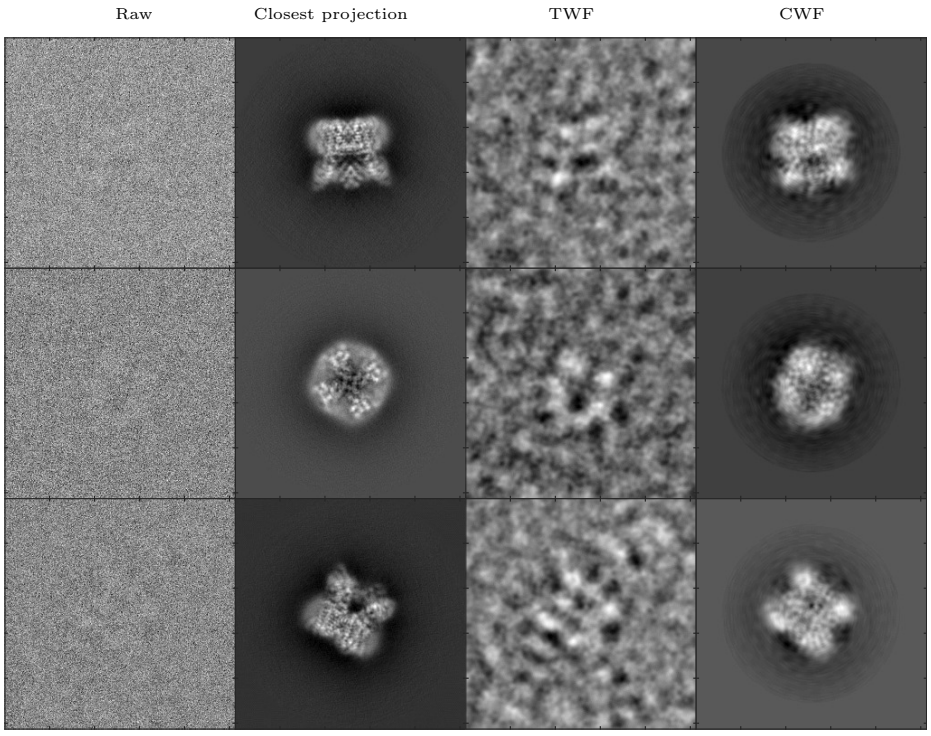


Figure 3.5: **Denoising an experimental dataset of TRPV1 [47]**: Here we show, for three images in the dataset, the raw image, the closest true projection image generated from the 3D reconstruction of the molecule (EMDB 5778), the denoised image obtained using TWF, and the denoised image obtained using CWF. In this experiment, 35645 images of size 256×256 belonging to 935 defocus groups were used. The amplitude contrast is 10%, the spherical aberration is 2mm, and the voltage is 300kV.

3.3.4 Experimental Dataset - 80S ribosome

We apply CWF to an experimental dataset of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine, taken using a FEI FALCON II 4k × 4k direct electron detector. The raw micrographs and picked particles are available on the public database EMPIAR as EMPIAR-10028, and the 3D reconstruction is available on EMDB as EMDB-2660, courtesy of Wong et al. [109]. The dataset we used was provided by Dr. Sjors Scheres, and consists of 105247 motion corrected, picked particle images of size 360 × 360 with a pixel size of 1.34$\mathring{A}$. Using 20 cores, calculating the Fourier Bessel coefficients took 731 seconds while covariance estimation and Wiener filtering together took 385 seconds. The result is shown in Figure 3.6. CWF retains 962 eigenvalues of Σ.

3.3.5 Experimental Dataset - IP_3R1

We apply CWF to an experimental dataset of the Inositol 1, 4, 5-triphosphate receptor 1 (IP_3R1) provided by Dr. Irina Serysheva, obtained using the older Gatan 4k × 4k CCD camera [49]. The 3D reconstruction obtained from this dataset is available on EMDB as EMDB-5278. The dataset consists of 37382 images of size 256×256 pixels with a pixel size of 1.81$\mathring{A}$. Using 20 cores, calculating the Fourier Bessel coefficients took 429 seconds while covariance estimation and Wiener filtering together took 589 seconds. The result is shown in Figure 3.7. CWF retains 290 eigenvalues of Σ.

3.3.6 Experimental Dataset - 70S ribosome

We apply CWF to an experimental dataset of the 70S ribosome provided by Dr. Joachim Frank's group [3]. This heterogeneous dataset consists of 216517 images of size 250 × 250 pixels with a pixel size of 1.5$\mathring{A}$, obtained using the older TVIPS TEMCAM-F415 (4k x 4k) CCD detector. The 3D reconstruction obtained from this

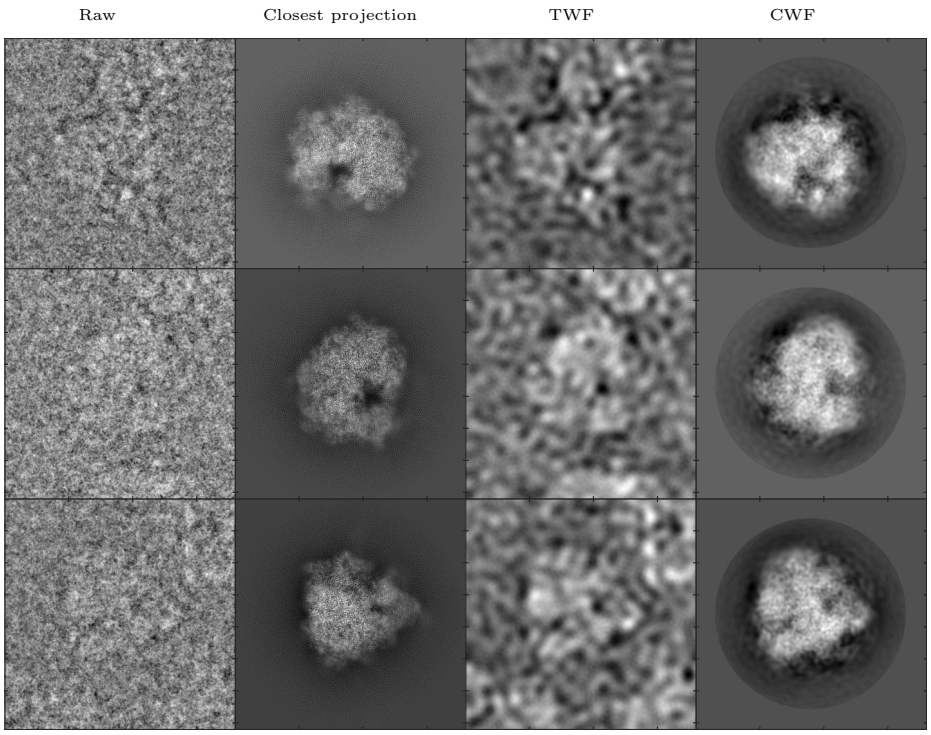


Figure 3.6: **Denoising an experimental dataset of the 80S ribosome [109]**: Here we show, for three images in the dataset, the raw image, the closest true projection image generated from the 3D reconstruction of the molecule (EMDB 2660), the denoised image obtained using TWF, and the denoised image obtained using CWF. In this experiment, the first 30000 images out of the 105247 images in the dataset were used for covariance estimation. The images are of size 360×360 and belong to 290 defocus groups. The amplitude contrast is 10%, the spherical aberration is 2mm, and the voltage is 300kV.

dataset is available on EMDB as EMDB-5360. Using 20 cores, calculating the Fourier Bessel coefficients took 1174 seconds while covariance estimation and Wiener filtering together took 113 seconds. The result is shown in Figure 3.8. CWF retains 219 eigenvalues of Σ.

3.3.7 Outlier Detection

In the cryo-EM pipeline, a significant amount of time is spent on discarding outliers by visual inspection after the particle picking step. CWF provides an automatic way to classify picked particles into "good" particles and outliers. The classifier uses the contrast of a denoised image to determine if it is an outlier.

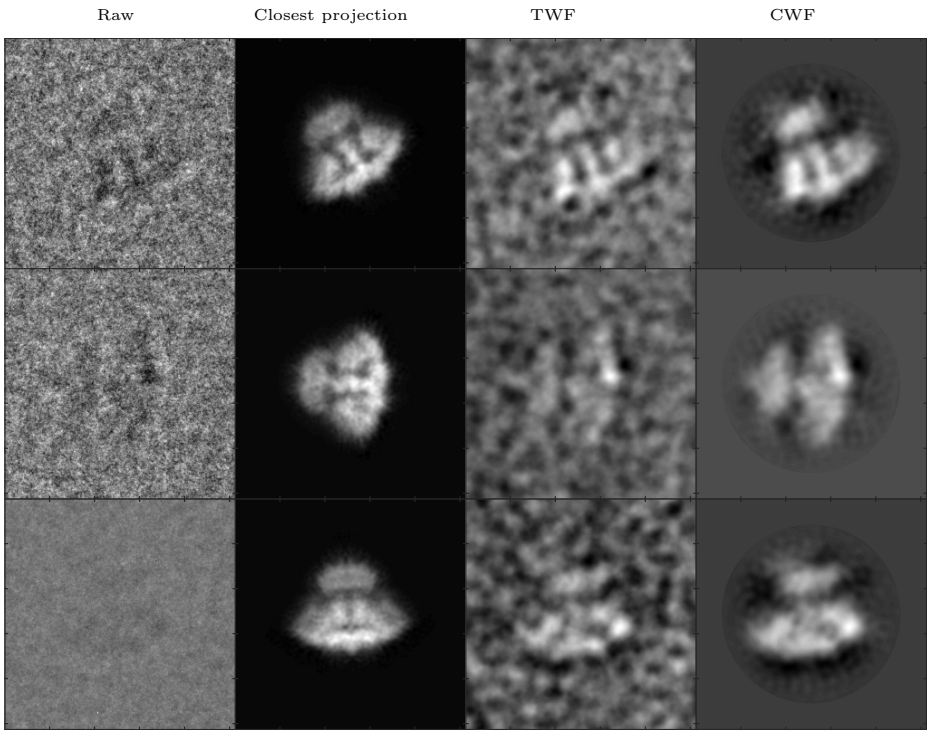


Figure 3.7: **Denoising an experimental dataset of IP_3R1 [49]**: Here we show, for three images in the dataset, the raw image, the closest true projection image generated from the 3D reconstruction of the molecule (EMDB 5278), the denoised image obtained using TWF, and the denoised image obtained using CWF. In this experiment, 37382 images of size 256×256 belonging to 851 defocus groups were used. The amplitude contrast is 15%, the spherical aberration is 2mm, and the voltage is 200kV.

The specimen particles can be at various depths in the ice layer at the time of imaging, so the acquired projection images can have different contrasts. The contrast can be modeled as an additional scalar parameter α for each acquired noisy projection image as in eqn. 3.19, typically as a uniformly distributed random variable spread about its mean at 1.

$$Y_i = \alpha_i A_i X_i + \xi_i, \quad i = 1, 2, \ldots, n \tag{3.19}$$

We absorb the contrast α into $\mathbf{X}$ and estimate $\alpha_i X_i$ in this case, using the same procedure as before. We perform an experiment with synthetic data generated using EMDB-6454 with additive colored Gaussian noise at SNR=1/20, and $\alpha \in [0.75, 1.5]$.

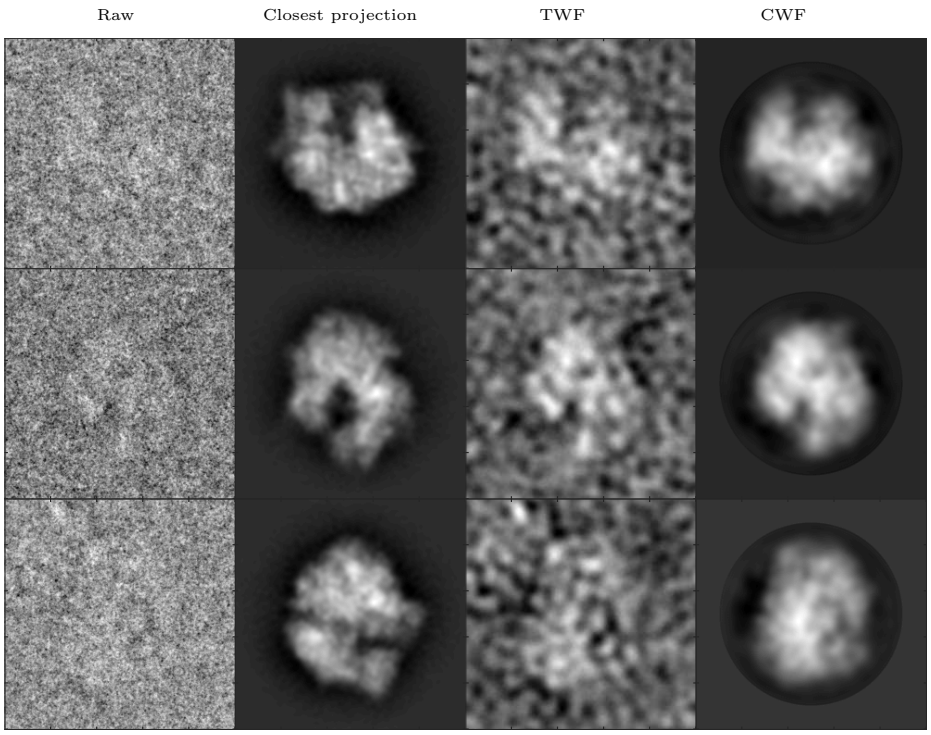


Figure 3.8: **Denoising an experimental dataset of 70S [3]**: Here we show, for three images in the dataset, the raw image, the closest true projection image generated from the 3D reconstruction of the molecule (EMDB 5360), the denoised image obtained using TWF, and the denoised image obtained using CWF. In this experiment, the first 99979 images out of the 216517 images in the dataset were used for covariance estimation. The images are of size 250×250 and belong to 38 defocus groups. The amplitude contrast is 10%, the spherical aberration is 2.26mm, and the voltage is 300kV.

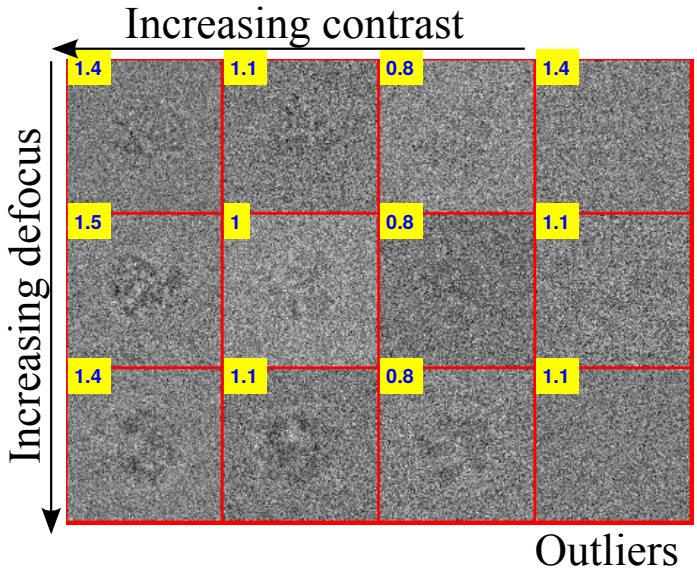
10% of the projection images are replaced by "outliers", that is, pure noise images containing no signal. Fig. 3.9c shows the estimated mean image μ, and Fig. 3.9d shows the top 6 principal components of the estimated covariance $\hat{\Sigma}$, also known as eigenimages. Fig. 3.9a and Fig. 3.9b show a sample of raw and denoised images respectively. High contrast images enjoy a higher SNR and are thus of interest for subsequent steps of the pipeline. On the other hand, outlier images, which typically have low contrast after denoising, can be automatically detected by a linear classifier after CWF and discarded from the dataset. In the experiment shown in Fig. 3.9a and 3.9b, a classifier with a threshold of 0.95 for the contrast discards 95% of the outliers, while 3% of the inliers are also discarded in the process.

One can also use a different classifier based on features like the relative energy of the image before and after denoising, etc. However, outliers that look like particles, for example, images belonging to a different class of a heterogeneous dataset which act as "contaminants", are difficult to detect using this method.

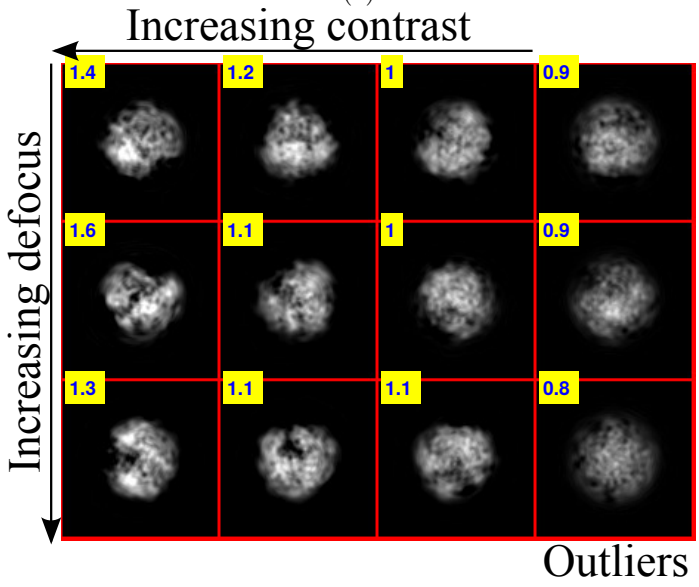
3.4 Conclusion

In this chapter we presented a new approach for image restoration of cryo-EM images, CWF, whose main algorithmic components are covariance estimation and deconvolution using Wiener filtering. CWF performs both CTF correction, by correcting the Fourier phases and amplitudes of the images, as well as denoising, by eliminating the noise thereby improving the SNR of the resulting images. In particular, since CWF applies Wiener filtering in the data-dependent basis of principal components ("eigenimages"), while TWF applies Wiener filtering in the data-independent Fourier basis, we see in numerical experiments that CWF performs better than TWF, and considerably better at high noise levels. We demonstrated the ability of CWF to restore images for several experimental datasets, acquired with both CCD detectors and the state-of-the-art direct electron detectors.

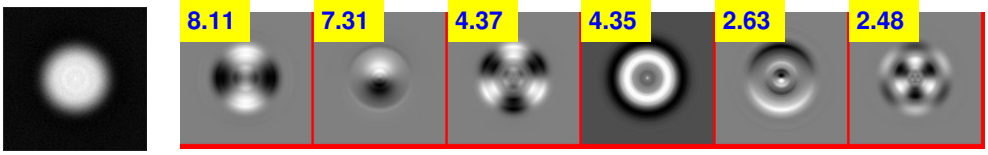
Due to the high noise level typical in cryo-EM images, 2D classification is performed before estimating a 3D ab-initio model. Class averages enjoy a higher SNR and are used to estimate viewing angles and obtain an initial model. For future work, it remains to be seen whether the resulting denoised images from CWF can be directly used to estimate viewing angles, without performing classification and averaging. Another possible future direction is integration of CWF into existing 2D class averaging procedures in order to improve their performance, which we elucidate in the next chapter.



(a)



(b)



(c) (d)

Figure 3.9: (a) **Raw images**: A sample of synthetic data generated using EMDB-6454 with additive colored Gaussian noise at SNR=1/20. 10% of the projection images are replaced by pure noise. The contrast parameter α ranges from 0.75 to 1.5. The outliers are shown in the last column. Inset in a yellow box is the contrast of each image. (b) **Denoised images**: The denoised images using CWF. Notice the low contrast outliers in the last column. (c) **Estimated Mean Image** (d) **Top 6 eigenimages**: Inset in a yellow box is the corresponding eigenvalue.

Chapter 4

Mahalanobis Distance for Class Averaging of Cryo-EM Images

4.1 Introduction

In SPR using cryo-EM, first, the sample, consisting of randomly oriented, nearly identical copies of a macromolecule, is frozen in a thin ice layer. An electron microscope is used to acquire top view images of the sample, in the form of a large image called a 'micrograph', from which individual particle images are picked semi-automatically. After preprocessing the selected raw particle images, the images are clustered. The images within each class are averaged to enjoy a higher SNR than the individual

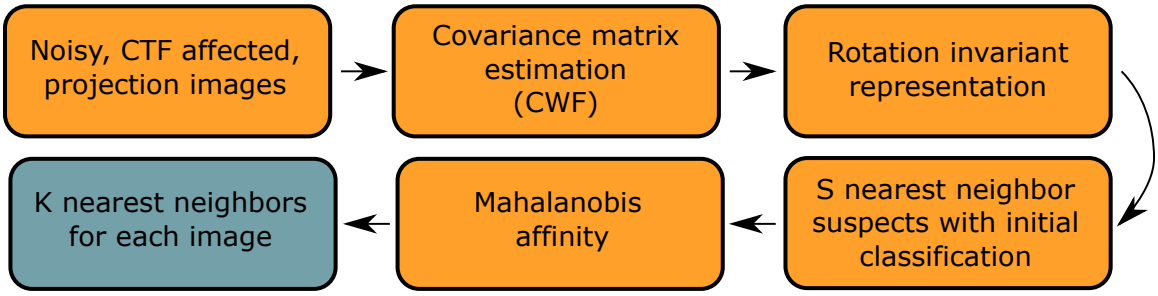


Figure 4.1: Pipeline of algorithm

images. This step is known as "class averaging". To minimize radiation damage, cryo-EM imaging must be constrained to low electron doses, which results in a very low SNR in the acquired 2D projection images. Class averaging is thus a crucial step in the SPR pipeline: class averages are used for a preliminary inspection of the dataset, to eliminate outliers, and in semi-automated particle picking [80]. Typically, a user manually picks particles from a small number of micrographs. These are used to compute class averages, which are further used as templates to pick particles from all micrographs. Class averages are also used in subsequent stages of the SPR pipeline, such as 3D ab-initio modeling.

The two popular approaches for 2D class averaging [67, 68, 79, 101, 62, 63] in cryo-EM are multivariate statistical analysis (MSA)[101] with multi-reference alignment (MRA) [19] and iterative reference-free alignment using K-means clustering [68]. Popular cryo-EM packages like RELION, XMIPP, EMAN2, SPIDER, SPARX, IMAGIC [103, 82, 96, 56, 80, 39] use some of these methods for class averaging. RELION uses a maximum likelihood classification procedure. A faster and more accurate approach for 2D class averaging based on rotationally invariant representation was introduced in [115] and is implemented in the cryo-EM software package ASPIRE (`http://spr.math.princeton.edu/`).

In chapter 3 [10], it was shown that preliminary inspection of the underlying clean images and outlier detection can be performed at an earlier stage, by better denoising the acquired images using an algorithm called Covariance Wiener Filtering (CWF). In CWF, the covariance matrix of the underlying clean projection images is estimated from their noisy, CTF-affected measurements. The covariance is then used in the classical Wiener deconvolution framework to obtain denoised images.

There are two main contributions of this chapter. First, we introduce a new similarity measure, related to the Mahalanobis distance [53], to compute the similarity of pairs of cryo-EM images. Second, we use the proposed Mahalanobis distance to

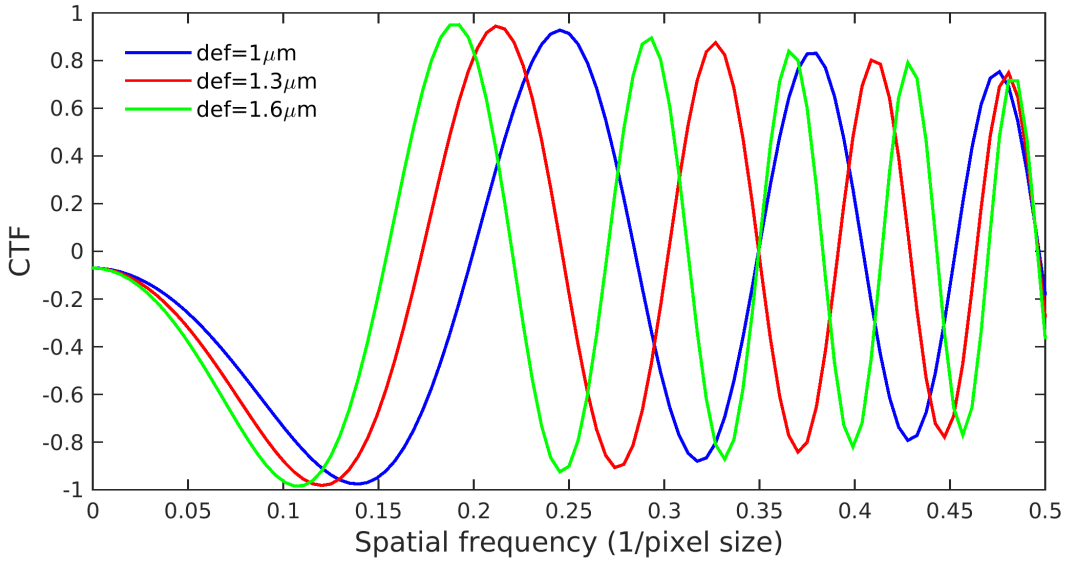


Figure 4.2: CTF's for different values of the defocus. CTF parameters used are: the amplitude contrast $\alpha = 0.07$, the electron wavelength $\lambda = 2.51pm$, the spherical aberration constant $Cs = 2.0$, the B-factor $B = 10$, the defocus= $1\mu m, 1.3\mu m$, and $1.6\mu m$, and the pixel size is $2.82\mathring{A}$. See eq. (4.3)

improve the class averaging algorithm described in [115]. We first obtain for each image a list of S other images suspected as nearest neighbors using the rotation invariant representation (see Sec. 4.2 for details), and then rank these suspects using the Mahalanobis distance (see Fig. 4.1). The top K nearest neighbors, where $K < S$, given by this procedure are finally aligned and averaged to produce class averages. We test the new algorithm on a synthetic dataset at various noise levels and observe an improvement in the number of nearest neighbors correctly detected.

4.2 Background

4.2.1 Image Formation Model

Under the linear, weak phase approximation (see [23, Chapter 2]), the image formation model in cryo-EM is given by

$$y_i = a_i \star x_i + n_i \tag{4.1}$$

where $\star$ denotes the convolution operation, y_i is the noisy projection image in real space, x_i is the underlying clean projection image in real space, a_i is the point spread function of the microscope, and n_i is additive Gaussian noise that corrupts the image. In the Fourier domain, images are multiplied with the Fourier transform of the point spread function, called the CTF, and eqn.(4.1) can be rewritten as

$$Y_i = A_i X_i + N_i \tag{4.2}$$

where Y_i, X_i and N_i are the Fourier transforms of y_i, x_i and n_i respectively. The CTF is approximately given by (see [23, Chapter 3])

$$CTF(\hat{k}; \Delta\hat{z}^2) = e^{-B\hat{k}^2} \sin[-\pi\Delta\hat{z}\hat{k}^2 + \frac{\pi}{2}\hat{k}^4] \tag{4.3}$$

where $\Delta\hat{z} = \frac{\Delta z}{[C_s\lambda]^{\frac{1}{2}}}$ is the "generalized defocus" and $\hat{k} = [C_s\lambda]^{\frac{1}{4}}k$ is the "generalized spatial frequency", and B is the B-factor for the Gaussian envelope function. CTF's corresponding to different defocus values have different zero crossings (see Fig.4.2). Note that the CTF inverts the sign of the image's Fourier coefficients when it is negative, and completely suppresses information at its zero crossings.

4.2.2 Rotationally Invariant Class Averaging

The procedure for class averaging, described in [115], was demonstrated to be both faster and more accurate than other existing class averaging procedures. It consists of three main steps. First, principal component analysis (PCA) of CTF-corrected phase flipped images is computed. We refer to this step as steerable PCA, because the procedure takes into account that the 2D covariance matrix commutes with in-plane rotations. Second, the bispectrum of the expansion coefficients in the reduced steerable basis is computed. The bispectrum is a rotationally invariant representation of images, but is typically of very high dimensionality. It is projected onto a lower dimensional subspace using a fast, randomized PCA algorithm [72]. One way to compare images after this step is using the normalized cross correlation. At low SNR, it is difficult to identify true nearest neighbors from the cross correlation. Therefore, Vector Diffusion Maps (VDM) [15] was used to further improve the initial classification by taking into account the consistency of alignment transformations among nearest neighbor suspects.

4.2.3 Covariance Wiener Filtering (CWF)

CWF was proposed in [10] (see chapter 3) as an algorithm to (i) estimate the CTF-corrected covariance matrix of the underlying clean images (since phase flipping is not an optimal correction) and (ii) using the estimated covariance to solve the associated deconvolution problem in eqn. 4.2 to obtain denoised images, that are estimates of X_i for each i in eqn. 4.2. The first step involves estimating the mean image of the dataset, μ, denoted $\hat{\mu}$, followed by solving a least squares problem to estimate the covariance Σ, denoted $\hat{\Sigma}$. Under the assumption of additive white Gaussian noise with variance σ^2, the estimate of the underlying clean image X_i is given by

$$\hat{X}_i = (I - H_i A_i)\hat{\mu} + H_i Y_i \tag{4.4}$$

where $H_i = \hat{\Sigma} A_i^T (A_i \hat{\Sigma} A_i^T + \sigma^2 I)^{-1}$

4.3 Anisotropic Affinity

The Mahalanobis distance in statistics [53] is a generalized, unitless and scale invariant similarity measure that takes correlations in the dataset into account. It is popularly used for anomaly detection and clustering [110, 112].

Our goal is to define a similarity measure to compare how close any two cryo-EM images are, given the CTF-affected, noisy observations for a pair of images from possibly different defocus groups, say Y_i and Y_j in eq. 4.2. CTF correction is a challenging problem due to the numerous zero crossings of the CTF. A popular, albeit, only partial correction of CTF is phase flipping, which involves simply inverting the sign of the Fourier coefficients. This corrects for the phase inversion caused by the CTF, but does not perform amplitude correction. Since phase flipping is suboptimal as a method for CTF correction, computing nearest neighbors using the Euclidean distance between features constructed from phase flipped, denoised images can suffer from incorrectly identified neighbors. One simple approach would be to use the Euclidean distance between the CWF denoised images, as a measure of similarity. However, the optimality criterion for obtaining CWF denoised images is different from that for identifying nearest neighbors. Also, after CWF denoising, noise is no longer white, so the Euclidean distance is not an optimal measure of affinity.

In our statistical model, the underlying clean images $X_1, X_2, \ldots X_n \in \mathbb{C}^d$ (where n is the total number of images and d is the total number of pixels in each image) are assumed to be independent, identically distributed (i.i.d.) samples drawn from a Gaussian distribution. Further, we assume that the noise in our model is additive white Gaussian noise

$$X_i \sim \mathcal{N}(\mu, \Sigma) \qquad N_i \sim \mathcal{N}(0, \sigma^2 I_d) \tag{4.5}$$

We note that while the assumption of a Gaussian distribution does not hold in practice, it facilitates the derivation of the new measure. The justification of the new measure is its superiority over the existing class averaging algorithm, as we demonstrate in Sec. 4.5.

The Gaussian assumption on signal and noise (5) and the image formation model (2) imply that Y_i is also Gaussian

$$Y_i \sim \mathcal{N}(A_i\mu, A_i\Sigma A_i^T + \sigma^2 I_d), \quad \text{for } i = 1, \ldots, n. \tag{4.6}$$

The joint distribution of (X_i, Y_i) is given by

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_i & I \end{bmatrix} \times \begin{bmatrix} X_i \\ N_i \end{bmatrix} \tag{4.7}$$

$$\sim \mathcal{N}\left(\begin{bmatrix} \mu \\ A_i\mu \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma A_i^T \\ A_i\Sigma & A_i\Sigma A_i^T + \sigma^2 I \end{bmatrix} \right) \tag{4.8}$$

The conditional distribution of X_i given Y_i is also Gaussian

$$X_i | Y_i = y_i \sim \mathcal{N}(\alpha_i, L_i) \tag{4.9}$$

where

$$\begin{aligned} \alpha_i &= \mu + \Sigma A_i^T (A_i\Sigma A_i^T + \sigma^2 I)^{-1}(y_i - A_i\mu) \\ L_i &= \Sigma - \Sigma A_i^T (A_i\Sigma A_i^T + \sigma^2 I)^{-1} A_i\Sigma. \end{aligned} \tag{4.10}$$

So

$$X_i - X_j | Y_i = y_i, Y_j = y_j \sim \mathcal{N}(\alpha_i - \alpha_j, L_i + L_j) \tag{4.11}$$

Let $X_i - X_j = x_{ij}$, and $\alpha_i - \alpha_j = \alpha_{ij}$. Then, for small ϵ, the probability that the ℓ_p

distance between X_i and X_j is smaller than ϵ is

$$\Pr(||X_{ij}||_p < \epsilon | Y_i = y_i, Y_j = y_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|L_i + L_j|^{\frac{1}{2}}} \times$$
$$\int_{B_p(0,\epsilon)} \exp\{-\frac{1}{2}(x_{ij} - \alpha_{ij})^T (L_i + L_j)^{-1}(x_{ij} - \alpha_{ij}\} dx_{ij} \tag{4.12}$$

$$= \frac{\epsilon^d \mathrm{Vol}(B_p(0,1))}{(2\pi)^{\frac{d}{2}}|L_i + L_j|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\alpha_{ij}^T (L_i + L_j)^{-1}\alpha_{ij}\} + \mathcal{O}(\epsilon^{d+1}) \tag{4.13}$$

where $B_p(0,1)$ is the ℓ_p ball of radius 1 in $\mathbb{R}^d$ centered at the origin (which follows from Taylor expanding the integrand). The probability of $\|X_{ij}\|_p < \epsilon$ given the noisy images y_i and y_j is a measure of the likelihood for the underlying clean images x_i and x_j to originate from the same viewing direction. We define our similarity measure after taking the logarithm on both sides of eqn. 4.13), dropping out the constants independent of i and j, and substituting back α_{ij}:

$$-\frac{1}{2}\log(|L_i + L_j|) - \frac{1}{2}(\alpha_i - \alpha_j)^T (L_i + L_j)^{-1}(\alpha_i - \alpha_j) \tag{4.14}$$

Notice the resemblance of the second term in eq. (4.14) to the classical Mahalanobis distance [53]. This term takes into account the anisotropic nature of the covariance matrix by appropriately normalizing each dimension when computing the distance between two points. Note that this distance is different for different pairs of points since it depends on $L_i + L_j$, unlike the Euclidean distance and the classical Mahalanobis distance.

SNR	This work	[115] (VDM)	[115] (No VDM)
1/40	58373	56896	49560
1/60	34965	32113	29219
1/100	17262	14431	13706

Table 4.1: Number of nearest neighbors with correlation > 0.9, using 10,000 images, $K = 10$ and $S = 50$.

4.4 Algorithm for Improved Class Averaging using Mahalanobis Distance

We propose an improved class averaging algorithm that incorporates the affinity measure (4.14). The quantities α_i, L_i are computed for each image and defocus group respectively (in practice Σ is replaced by its estimate $\hat{\Sigma}$), using CWF [10]. The estimated covariance using CWF is block diagonal in the Fourier Bessel basis. In practice, we use α_i, L_i projected onto the subspace spanned by the principal components (for each angular frequency block). We obtain an initial list of S nearest neighbors for each image using the Initial Classification algorithm in [115]. Then, for the list of nearest neighbors corresponding to each image, the affinity (4.14) is computed and used to pick the closest K nearest neighbors, where $K < S$.

4.5 Numerical experiments

We test the improved class averaging algorithm on a synthetic dataset that consists of projection images generated from the volume of P. falciparum 80S ribosome bound to E-tRNA, available on the Electron Microscopy Data Bank (EMDB) as EMDB 6454. The algorithm was implemented in the UNIX environment, on a machine with total RAM of 1.5 TB, running at 2.3 GHz, and with 60 cores. For the results described here, we used $10,000$ projection images of size 65×65 that were affected by the various CTF's and additive white Gaussian noise at various noise levels, in particular, we

show here results for 4 values of the SNR. The images were divided into 20 defocus groups. Initial classification was first used to select $S = 50$ nearest neighbors for each image. After rotationally aligning the suspected neighbors, the Mahalanobis distance was computed between each image and the 50 aligned suspects. We then pick the closest $K = 10$ neighbors for each image (in practice, the choice of K depends on the SNR and the number of images). For comparison, we compute 10 nearest neighbors for each image using only Initial Classification (with or without using the optional VDM step). Table 4.1 shows the number of pairs of nearest neighbor images detected with each method at various SNR's, that have correlation > 0.9 between the original clean images, indicating that they are indeed neighbors. We note an improvement in the number of true nearest neighbors detected by the improved classification algorithm using the Mahalanobis distance. Figure 4.3 shows the estimated probability density function of the angular distance between nearest neighbor images, using 1) Initial Classification only 2) Improved classification using the Mahalanobis distance by repeating this experiment at four different SNR's. Figure 4.4a shows the results of Initial Classification and the improved class averaging algorithm on this synthetic dataset. We compare the quality of the class averages from [115] and this paper with $K = 10$. Averaging over a large number of nearest neighbors reduces the noise variance. However, it also blurs the underlying clean signal, since the neighbors are not exactly from the same viewing direction. Therefore, it is crucial to correctly identify only the top few nearest neighbors and average them in order to sufficiently reduce the noise without blurring the features too much. We see in Figure 4.4b that noise reduces in the class averages when K increases. The procedure in [115] took 168 seconds, while the improved classification using the anisotropic affinity took 860 seconds for the experiment described here.

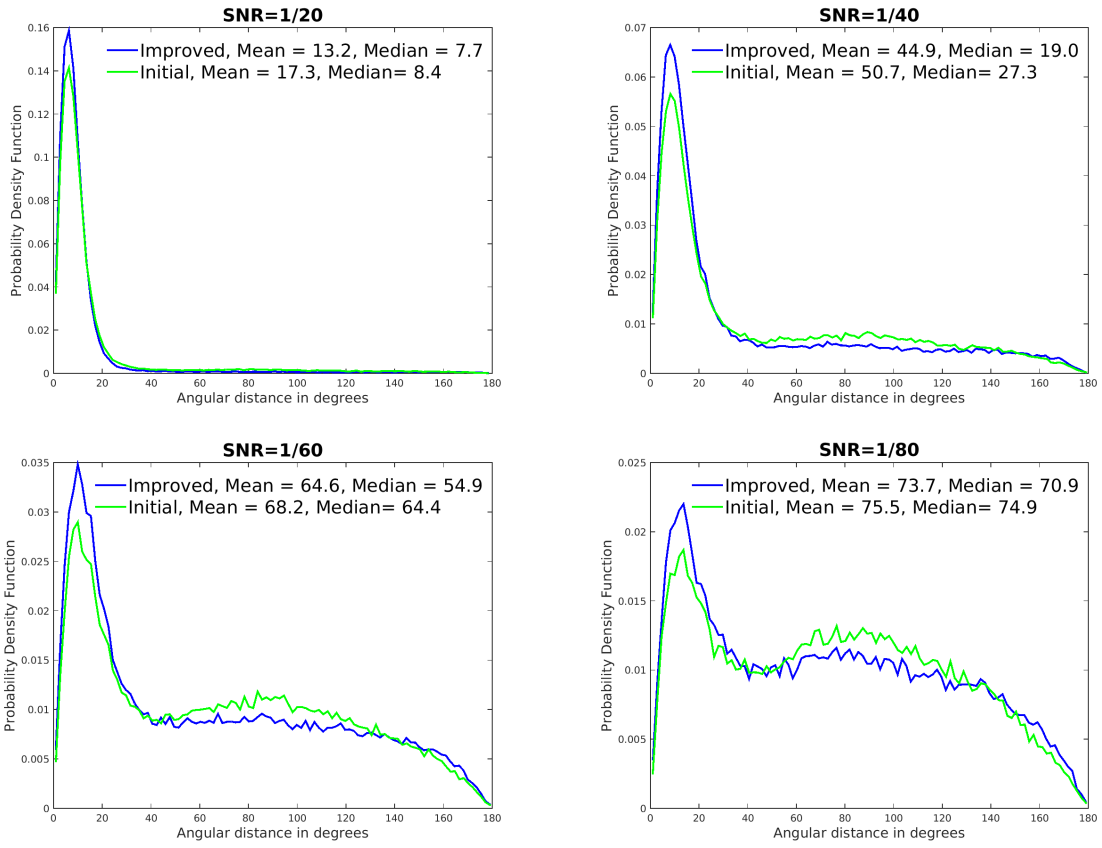
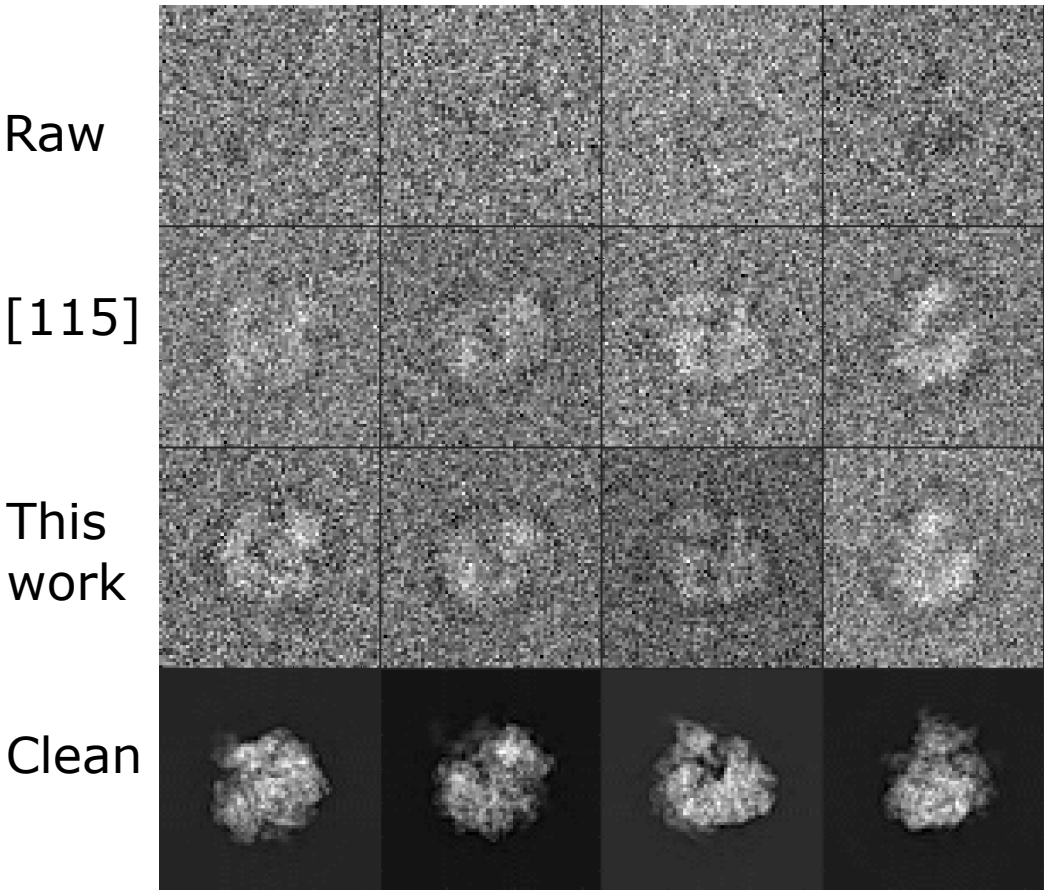
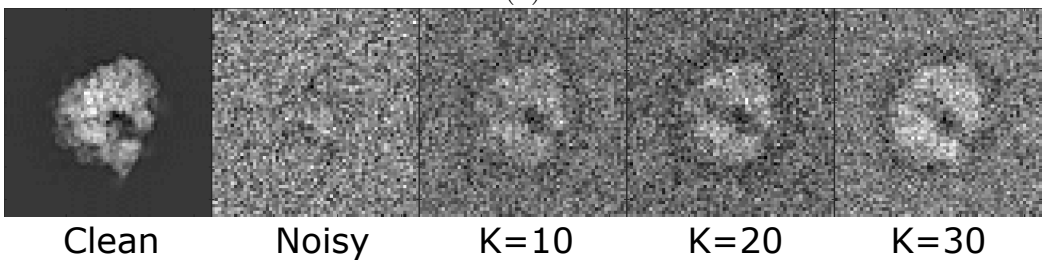


Figure 4.3: The estimated probability density function of the angular distance (in degrees) between images classified into the same class by 1) Initial Classification and 2) Improved Classification using the anisotropic affinity at different SNR's.



(a)



(b)

Figure 4.4: Results of class averaging of a synthetic dataset of 10000 projection images of size 65×65, affected by CTF and SNR $= 1/40$. (a) We show class averages with Initial Classification in the second row, and with the improved algorithm using the anisotropic affinity in the third row. We use $K = 10$ and $S = 50$. (b) Class averages for one image in the dataset with the improved algorithm using the anisotropic affinity, for $S = 50$, and using $K = 10,\ 20,\ 30$.

4.6 Discussion

We introduced a new similarity measure to compare CTF-affected cryo-EM images belonging to different defocus groups. The anisotropic affinity derived in this paper is similar to the one that appears in [86, 95] but also includes an additional logarithmic term. We provided a new probabilistic interpretation for this anisotropic affinity. The affinity can also be used as a similarity measure for any manifold learning procedure [95, 86] such as diffusion maps [89, 15], with or without missing data, and extended to other imaging modalities where images are affected by different point spread functions or blurring kernels.

Chapter 5

Anisotropic twicing for single particle reconstruction using autocorrelation analysis

5.1 Introduction

The missing phase problem in crystallography entails recovering information about a crystal structure that is lost during the process of imaging. In X-ray crystallography, the measured diffraction patterns provide information about the modulus of the 3D Fourier transform of the crystal. The phases of the Fourier coefficients need to be recovered by other means, in order to reconstruct the 3D electron density map of the crystal. A popular method to solve the missing phase problem is Molecular Replacement (MR) [74, 73, 78], which relies on a previously solved homologous structure which is similar to the unknown structure. The unknown structure is then estimated using the Fourier magnitudes of its diffraction data, along with phases from the homologous structure.

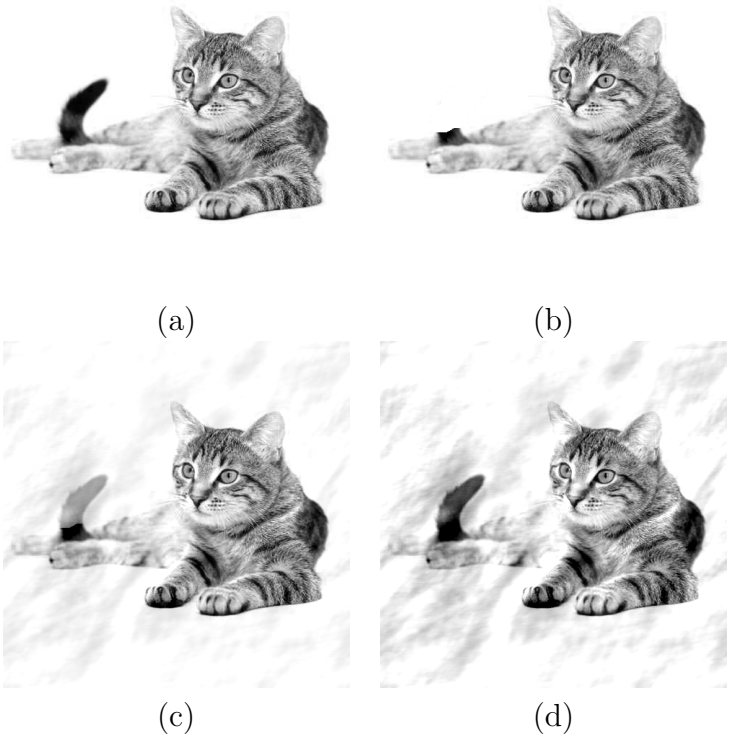
The missing phase problem can be formulated mathematically using matrix nota-

tion that enables generalization as follows. Each Fourier coefficient $\boldsymbol{A}$ is a complex-valued scalar, i.e., $\boldsymbol{A} \in \mathbb{C}^{1\times 1}$ that we wish to estimate, given measurements of $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^*$ ($\boldsymbol{A}^*$ denotes the complex conjugate transpose of $\boldsymbol{A}$, i.e., $\boldsymbol{A}^*_{ij} = \overline{A_{ji}}$), corresponding to the Fourier squared magnitudes, and $\boldsymbol{B}$ corresponds to a previously solved homologous structure such that $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{E}$, where $\boldsymbol{E}$ is a small perturbation. We denote an estimator of $\boldsymbol{A}$ as $\hat{\boldsymbol{A}}$. There are many possible choices for such an estimator. One such choice is the solution to the least squares problem

$$\hat{\boldsymbol{A}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{A} - \boldsymbol{B}\|_F, \text{ subject to } \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{C} \tag{5.1}$$

where $||.||_F$ denotes the Frobenius norm. However, it has been noticed that $\hat{\boldsymbol{A}}_{\mathrm{LS}}$ does not reveal the correct relative magnitude of the unknown part of the crystal structure, and the recovered magnitude is about half of the actual value. As a magnitude correction scheme, it was empirically found that setting the magnitude to be twice the experimentally measured magnitude minus the magnitude of the homologous structure has the desired effect of approximately resolving the issue. That is, the estimator $2\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}$ is used instead. The theoretical advantage of this unbiased estimator for the case when $\boldsymbol{A} \in \mathbb{C}^{1\times 1}$ has been justified in [54]. Following [97], we refer to this procedure as twicing.

The advantage of using twicing is demonstrated in the following illustrative toy experiment [16] for the 2D case. We start with an image of a cat with a tail, which is the unknown image that we want to recover. We are given the Fourier magnitudes of the unknown image, measured in an experiment. In analogy with a known homologous structure used in MR, we have access to a similar image, that of a cat, but with its tail missing. We show the results of retrieving the original image using least squares, with and without employing twicing for magnitude correction, and note that twicing restores the tail better than least squares (see Fig 5.1).



(a) (b)

(c) (d)

Figure 5.1: Demonstrating twicing in MR through a toy example [16]: given an unknown image whose Fourier magnitudes are known through measurements, but phases are missing, and a known similar image for which both the Fourier magnitudes and phases are completely known. (a) Original image: unknown phases, known magnitudes (b) Similar image: known phases and magnitudes (note that the tail is missing) (c) Least squares estimator of original image, no magnitude correction (d) Twicing for magnitude correction. Note that the tail is better restored when twicing is used.

As a natural generalization, one might wonder whether the estimator $2\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}$ performs well for the non-scalar case $\boldsymbol{A} \in \mathbb{R}^{N\times D}$ (or $\mathbb{C}^{N\times D}$), where $(N, D) \neq (1, 1)$. In this paper, we consider the following problem: How to estimate $\boldsymbol{A} \in \mathbb{R}^{N\times D}$ (or $\mathbb{C}^{N\times D}$) from $\boldsymbol{C}$ and $\boldsymbol{B}$, where $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^*$ and $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{E}$ for a matrix $\boldsymbol{E}$ of small magnitude? When $N = D$, the result derived in this paper (see Sec. 5.4) for an "asymptotically consistent" estimator of $\boldsymbol{A}$ is given by $\hat{\boldsymbol{A}}_{\mathrm{AT}} = \boldsymbol{B} + \boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^{\star}(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B})$ where $\boldsymbol{U}$ and $\boldsymbol{W}$ are defined in Sec. 5.4. In particular, when $\boldsymbol{A} \in \mathbb{C}^{1\times 1}$, this result coincides with the result in [54] and justifies the approach of twicing. A formal proof of the result derived in this paper is provided later in Sec. 5.5.

The motivation to study this problem is 3D structure determination in single particle reconstruction (SPR) without estimating the viewing angle associated with each image. Although we focus on cryo-electron microscopy (cryo-EM) here, the methods in this paper can also be applied to SPR using X-ray free electron lasers (XFEL). In SPR using XFEL, short but intense pulses of X-rays are scattered from the molecule. The measured 2D diffraction patterns in random orientations are used to reconstruct the 3D diffraction volume by an iterative refinement procedure, akin to the approach in cryo-EM. Recently, there have been attempts to use Kam's theory for SPR using XFEL, to determine the 3D diffraction volume without any iterative refinement [75, 32, 93].

In this paper, we revisit Orthogonal Extension (OE) in cryo-EM [9] that combines ideas from MR and Kam's autocorrelation analysis [35, 36] for the purpose of 3D homology modeling, that is, for reconstruction of an unknown complex directly from its raw, noisy images when a previously solved similar complex exists. In SPR using cryo-EM [44, 5, 30], the 3D structure of a macromolecule is reconstructed from its noisy, contrast transfer function (CTF) affected 2D projection images. Individual particle images are picked from micrographs, preprocessed, and used in further parts of the cryo-EM pipeline to obtain the 3D density map of the macromolecule. There

exist many algorithms in popular cryo-EM software such as RELION, XMIPP, SPIDER, EMAN2, FREALIGN [80, 56, 82, 96, 27] that, given a starting 3D structure, refine it using the noisy 2D projection images. The result of the refinement procedure is often dependent on the choice of the initial model. It is therefore important to have a procedure to provide a good starting model for refinement. Also, a high quality starting model may significantly reduce the computational time associated with the refinement procedure (although we note recent advances in fast refinement [8, 71]). Such a high quality starting model can be obtained using OE. The main computational component of autocorrelation analysis is estimation of the covariance matrix of the 2D images. This computation requires only a single pass over the experimental images [113, 10]. Autocorrelation analysis is therefore much faster than iterative refinement, which typically takes many iterations to converge. In fact, the computational cost of autocorrelation analysis is even lower than that of a single refinement iteration, as the latter involves comparison of image pairs (noisy raw images with volume projections). OE can also be used for the purpose of model validation, being a complementary method for structure prediction.

There are a few existing methods for ab-initio modeling. The random conical tilt method [51] can be used when two electron micrographs, one tilted and one untilted, are acquired with the same field of view. There are two main approaches for ab-initio estimation that do not involve tilting. One approach is to use the method of moments, that leverages the second order moments of the unknown 3D volume to estimate the particle orientations, but it suffers from being very sensitive to errors in the data [76, 25]. The other approach is based on using common-lines between images [29, 98, 87, 88]. However, common-lines based approaches have not been successful in obtaining 3D ab-initio models directly from raw, noisy images without performing any class averaging to suppress the noise.

OE predicts the structure directly from the raw, noisy images without any av-

eraging. The method is analogous to MR in X-ray crystallography for solving the missing phase problem. In OE, the homologous structure is used for estimating the missing orthogonal matrices associated with the spherical harmonics expansion of the 3D structure in reciprocal space. It is important to note that the missing orthogonal matrices in OE are not associated with the unknown pose of the particles, but with the spherical harmonics expansion coefficients. The missing coefficient matrices are, in general, rectangular of size $N \times D$, which serves as the motivation to extend twicing to the general case of finding an estimator when $(N, D) \neq (1, 1)$.

The paper is organized as follows: First, we briefly review Kam's theory for autocorrelation analysis and describe the problem of OE in cryo-EM in Sec. 5.2. Next, in Sec. 5.3, we describe the least squares solution to find an estimator to an unknown structure when we have noisy projection images of the unknown structure, and additional information about a homologous structure. In Sec. 4, we introduce Anisotropic Twicing as well as a family of estimators that interpolate between the least squares estimator and Anisotropic Twicing. We detail the procedure to estimate autocorrelation matrices and the algorithm of Orthogonal Extension with the Anisotropic Twicing estimator in Sec. 5. We benchmark the performance of these estimators through numerical experiments with synthetic and experimental datasets in Sec. 6. We provide a formal proof for asymptotic consistency of our Anisotropic Twicing correction scheme in the general case of $(N, D) \neq (1, 1)$ in the appendix (see Sec. 5.5). The code for all the algorithms in this paper is available in the open source software toolbox, ASPIRE, available for download at `spr.math.princeton.edu`.

We apply anisotropic twicing to both synthetic and experimental cryo-EM datasets, and find that it recovers the unknown structure better than the least squares and twicing estimators on synthetic data. This is the first demonstration of reconstructing a starting 3D model in the presence of experimental conditions of CTF and noise without any class averaging, directly from raw images using the 'Orthogonal Extension'

procedure [9]. While the anisotropic twicing estimator outperforms other estimators on synthetic datasets, in the case of the experimental dataset the reconstructions from all estimators are similar in quality, and any of these reconstructions can be used as a good starting point for refinement.

5.2 Orthogonal Extension (OE) in Cryo-EM

In [9], the authors presented two new approaches, collectively termed 'Orthogonal Retrieval' methods, for 3D homology modeling based on Kam's theory [35]. Orthogonal Retrieval can be regarded as a generalization of the MR method from X-ray crystallography to cryo-EM.

Let $\Phi_A : \mathbb{R}^3 \to \mathbb{R}$ be the electron scattering density of the unknown structure, and let $\mathcal{F}(\Phi_A) : \mathbb{R}^3 \to \mathbb{C}$ be its 3D Fourier transform. Consider the spherical harmonics expansion of $\mathcal{F}(\Phi_A)$

$$\mathcal{F}(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(k) Y_l^m(\theta, \varphi) \tag{5.2}$$

where k is the radial frequency and Y_l^m are the real spherical harmonics. Kam showed that the autocorrelation matrices

$$C_l(k_1, k_2) = \sum_{m=-l}^{l} A_{lm}(k_1)\overline{A_{lm}(k_2)}, \quad l = 0, 1, \ldots \tag{5.3}$$

can be estimated from the covariance matrix of the 2D projection images whose viewing angles are uniformly distributed over the sphere. This can be achieved with both clean as well as noisy images, as long as the number of noisy images is large enough to allow estimation of the underlying population covariance matrix of the clean images to the desired level of accuracy, using [10]. The decomposition (3) suggests that the l'th order autocorrelation matrix $\boldsymbol{C_l}$ has a maximum rank of $2l + 1$,

and the maximum rank is even smaller in the presence of symmetry.

While (5.2) is true if we want to represent the molecule to infinitely high resolution, in practice the images are sampled on a finite pixel grid and we cannot recover information beyond the Nyquist frequency. In addition, the molecule is compactly supported in $\mathbb{R}^3$, and the support size can also be estimated from the images. It is therefore natural to expand the volume in a truncated basis of spherical Bessel functions or 3D prolates. This leads to

$$\mathcal{F}(\Phi_A)(k,\theta,\varphi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} A_{lm}(k) Y_l^m(\theta,\varphi), \quad l = 0, 1, \ldots, L \tag{5.4}$$

where the truncation L is based on the resolution limit that can be achieved by the reconstruction. Our specific choice of L is described after (5.8). We can expand $A_{lm}(k)$ in a truncated basis of radial functions, chosen here as the spherical Bessel functions, as follows:

$$A_{lm}(k) = \sum_{s=1}^{S_l} a_{lms} j_{ls}(k). \tag{5.5}$$

Here the normalized spherical Bessel functions are

$$j_{ls}(k) = \frac{1}{c\sqrt{\pi}|j_{l+1}(R_{l,s})|} j_l(R_{l,s}\frac{k}{c}), \quad 0 < k < c, \quad s = 1, 2, \ldots, S_l, \tag{5.6}$$

where c is the bandlimit of the images, and $R_{l,s}$ is the s'th positive root of the equation $j_l(x) = 0$. The functions j_{ls} are normalized such that

$$\int_0^c j_{ls}(k) j_{ls}^*(k) k^2 dk = 1 \tag{5.7}$$

The number of radial basis functions S_l in (5.5) is determined using the Nyquist criterion, similar to [41, 113], where it has been described for 2D images expanded in a Fourier-Bessel basis (rather than 3D volumes as done here). We assume that the 2D images, and hence the 3D volume, are compactly supported on a disk of radius R and

have a bandlimit $0 < c \leq 0.5$. We require that the maximum of the inverse Fourier transform of the spherical Bessel function and its first zero after this maximum are both inside the sphere of compact support radius R. The truncation limit S_l in (5.5) is then defined by the sampling criterion as the largest integer s that satisfies [14]

$$R_{l,(s+1)} \leq 2\pi cR. \tag{5.8}$$

L in (5.4) is the largest integer l for which (5.8) has only one solution, that is, S_l in (5.5) is at least 1. Each $\boldsymbol{C_l}$ is a matrix of size $S_l \times S_l$ when using the representation (5.5) in (5.3). S_l is a monotonically decreasing function of l with approximately linear decay that we compute numerically. In matrix notation, (5.3) can be written as

$$\boldsymbol{C_l} = \boldsymbol{A_l}\boldsymbol{A_l^*}, \tag{5.9}$$

where $\boldsymbol{A_l}$ is a matrix of size $S_l \times (2l+1)$, with $A_l(s,m) = a_{lms}$ in (5.5). From (5.9), we note that $\boldsymbol{A_l}$ can be obtained from the Cholesky decomposition of $\boldsymbol{C_l}$ up to a unitary matrix $\boldsymbol{U_l} \in \mathrm{U}(2l+1)$ (the group of unitary matrices of size $(2l+1) \times (2l+1)$). Since Φ_A is real-valued, one can show using properties of its Fourier transform together with properties of the real spherical harmonics, that $A_{lm}(k)$ (and hence $\boldsymbol{A_l}$) is real for even l and purely imaginary for odd l. So $\boldsymbol{A_l}$ is unique up to an orthogonal matrix $\boldsymbol{O_l} \in \mathrm{O}(2l+1)$ (the group of orthogonal matrices of size $(2l+1) \times (2l+1)$). Determining $\boldsymbol{O_l}$ is the orthogonal retrieval problem in [9].

If $S_l > 2l+1$, estimating the missing orthogonal matrix $\boldsymbol{O_l}$ is equivalent to estimating $\boldsymbol{A_l}$. Since S_l is a decreasing function of l, for some large enough l we would have $S_l < 2l+1$. For example, for the largest $l = L$ where $S_L = 1$, $\boldsymbol{A_L}$ is of size $1 \times (2L+1)$, that is, it has $O(L)$ degrees of freedom. In such cases it does not make sense to estimate $\boldsymbol{O_L}$ which has $O(L^2)$ degrees of freedom. But we can still estimate $\boldsymbol{A_L}$ closest to $\boldsymbol{B_L}$ using (5.1).

5.3 The Least Squares Estimator

In this section we review the least squares estimator that was proposed in [9]. In order to determine the 3D Fourier transform $\mathcal{F}(\Phi_A)$ and thereby the 3D density Φ_A, we need to determine the coefficient matrices $\boldsymbol{A_l}$ of the spherical harmonic expansion. In OE, the coefficient matrices $\boldsymbol{A_l}$ are estimated with the aid of a homologous structure Φ_B. Suppose Φ_B is a known homologous structure, whose 3D Fourier transform $\mathcal{F}(\Phi_B)$ has the following spherical harmonic expansion:

$$\mathcal{F}(\Phi_B)(k,\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} B_{lm}(k) Y_l^m(\theta,\varphi) \tag{5.10}$$

In practice, the homologous structure Φ_B is available at some finite resolution, therefore only a finite number of coefficient matrices $\boldsymbol{B_l}$ $(l = 0, 1, \ldots, L_B)$ are given. We show how to estimate the unknown structure Φ_A up to the resolution dictated by the input images and the resolution of the homologous structure through estimating the coefficient matrices $\boldsymbol{A_l}$ for $l = 0, 1, \ldots, L_A$ where $L_A = \min(L, L_B)$.

Let $\boldsymbol{F_l}$ be any matrix of size $S_l \times 2l+1$ satisfying $\boldsymbol{C_l} = \boldsymbol{F_l}\boldsymbol{F_l}^*$, determined from the Cholesky decomposition of $\boldsymbol{C_l}$. Then, using (5.9)

$$\boldsymbol{A_l} = \boldsymbol{F_l}\boldsymbol{O_l} \tag{5.11}$$

where $\boldsymbol{O_l} \in \mathrm{O}(2l+1)$ (for $S_l > 2l+1$). Using the assumption that the structures are homologous, $\boldsymbol{A_l} \approx \boldsymbol{B_l}$, one can determine $\boldsymbol{O_l}$ as the solution to the least squares problem

$$\boldsymbol{O_l} = \underset{\boldsymbol{O}\in\mathrm{O}(2l+1)}{\arg\min} \|\boldsymbol{F_l}\boldsymbol{O} - \boldsymbol{B_l}\|_F^2, \tag{5.12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Although the orthogonal group is non-

convex, there is a closed form solution to (5.12) (see, e.g., [38]) given by

$$\boldsymbol{O}_l = \boldsymbol{V}_l \boldsymbol{U}_l^T, \tag{5.13}$$

where

$$\boldsymbol{B}_l^* \boldsymbol{F}_l = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{V}_l^T \tag{5.14}$$

is the singular value decomposition (SVD) of $\boldsymbol{B}_l^* \boldsymbol{F}_l$. Thus, $\boldsymbol{A}_l$ can be estimated by the following least squares estimator:

$$\hat{\boldsymbol{A}}_{l,\mathrm{LS}} = \boldsymbol{F}_l \boldsymbol{V}_l \boldsymbol{U}_l^T. \tag{5.15}$$

Hereafter, we drop the subscript l for convenience, since the procedure can be applied to each l separately.

5.3.1 Algorithm 1: Orthogonal Extension by Least Squares

Algorithm 1 Orthogonal Extension

procedure ORTHOGONAL EXTENSION BY LEAST SQUARES (OE-LS): ESTIMATE $\boldsymbol{A}$ GIVEN $\boldsymbol{B} \approx \boldsymbol{A}$, SUBJECT TO $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^*$

Input: $\boldsymbol{B} \in \mathbb{C}^{N \times D}$, $\boldsymbol{C} \in \mathbb{C}^{N \times N}$

Cholesky decomposition of $\boldsymbol{C}$ to find an $\boldsymbol{F} \in \mathbb{C}^{N \times D}$ such that $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^*$

Calculate $\boldsymbol{B}^*\boldsymbol{F}$ and its singular value decomposition $\boldsymbol{B}^*\boldsymbol{F} = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^*$.

The estimator is $\hat{\boldsymbol{A}}_{\mathrm{LS}} = \boldsymbol{F}\boldsymbol{V}_0\boldsymbol{U}_0^*$.

5.4 Unbiased Estimator: Anisotropic Twicing

The case that $\boldsymbol{A}$ is a complex-valued scalar, i.e., $\boldsymbol{A} \in \mathbb{C}^{1\times 1}$ has been studied in X-ray crystallography. The theoretical advantage of the unbiased estimator $2\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}$ for this case was elucidated in [54]. As a natural generalization, one may wonder whether

the estimator $2\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}$ is also unbiased for $(N, D) \neq (1, 1)$. We assume that $\boldsymbol{A}$ is sampled from the model $\boldsymbol{A} = \boldsymbol{F}\boldsymbol{V}$, where $\boldsymbol{F}\boldsymbol{F}^* = \boldsymbol{C}$ and $\boldsymbol{V}$ is a random orthogonal matrix (or a random unitary matrix) sampled from the uniform distribution with Haar measure over the orthogonal group when $\boldsymbol{A}$ is a real-valued matrix or the unitary group (when $\boldsymbol{A}$ is a complex-valued matrix). This probabilistic model is reasonable for (5.1), because when $\boldsymbol{A}\boldsymbol{A}^\star$ is given, $\boldsymbol{F}$ is known and $\boldsymbol{V}$ is an unknown orthogonal or unitary matrix, that is, we have no prior information about $\boldsymbol{V}$. In addition, we assume that $\boldsymbol{B}$ is a matrix close to $\boldsymbol{A}$ such that $\boldsymbol{A} - \boldsymbol{B}$ is fixed. Our goal is to find an unbiased estimator of $\boldsymbol{A}$ which is an affine transformation of $\hat{\boldsymbol{A}}_{\mathrm{LS}}$. The main result is as follows:

Theorem 5.4.1. *When $N = D$, assuming that the spectral decomposition of $\boldsymbol{C}$ is given by $\boldsymbol{C} = \boldsymbol{U}\,\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_D)\boldsymbol{U}^*$, then using our probabilistic model we have*

$$\mathbb{E}[\boldsymbol{A} - \hat{\boldsymbol{A}}_{LS}] = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*(\boldsymbol{A} - \boldsymbol{B}) + o(\|\boldsymbol{A} - \boldsymbol{B}\|_F), \tag{5.16}$$

where $\boldsymbol{T}$ is a diagonal matrix with i-th diagonal entry given by

$$\boldsymbol{T}_{ii} = \begin{cases} \frac{1}{D}\Big[-\frac{1}{2} + \sum_{1\leq j\leq D} \frac{\lambda_i^2}{\lambda_i^2+\lambda_j^2}\Big] & \text{when } \boldsymbol{A}, \boldsymbol{C} \in \mathbb{R}^{D\times D}, \\ \frac{1}{D}\sum_{1\leq j\leq D} \frac{\lambda_i^2}{\lambda_i^2+\lambda_j^2} & \text{when } \boldsymbol{A}, \boldsymbol{C} \in \mathbb{C}^{D\times D}, \end{cases}$$

and $f(\boldsymbol{X}) = o(\|\boldsymbol{X}\|_F)$ means that $\limsup_{\|\boldsymbol{X}\|_F\to 0} f(\boldsymbol{X})/\|\boldsymbol{X}\|_F \to 0$.

From (5.16), we have

$$(-\mathbf{I} + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*)(\boldsymbol{A} - \boldsymbol{B}) = \boldsymbol{B} - \mathbb{E}[\hat{\boldsymbol{A}}_{\mathrm{LS}}] + o(\|\boldsymbol{A} - \boldsymbol{B}\|_F)$$

and an "asymptotically consistent" estimator of $\boldsymbol{A}$ is given by

$$\hat{\boldsymbol{A}}_{\mathrm{AT}} = \boldsymbol{B} - (\mathbf{I} - \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*)^{-1}(\boldsymbol{B} - \hat{\boldsymbol{A}}_{\mathrm{LS}}) = \boldsymbol{B} + \boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^*(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}), \tag{5.17}$$

where $\boldsymbol{W} = (\mathbf{I} - \boldsymbol{T})^{-1}$.

A formal proof of Theorem 5.4.1 is provided in the appendix (Sec. 5.5). In particular, when $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{1\times 1}$, the matrices reduce to scalars: $\boldsymbol{U} = 1$, $\boldsymbol{T} = \frac{1}{2}$, $\boldsymbol{W} = 2$ and $\hat{\boldsymbol{A}}_{\mathrm{AT}} = \boldsymbol{B} + 2(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B})$. This result coincides with the result in [54] and justifies the approach of "twicing".

5.4.1 A family of estimators

From (5.16) it follows that

$$\mathbb{E}[\boldsymbol{A}] = \mathbb{E}[\hat{\boldsymbol{A}}_{\mathrm{LS}}] + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*(\boldsymbol{A} - \boldsymbol{B}) + o(\|\boldsymbol{A} - \boldsymbol{B}\|_F). \tag{5.18}$$

Following the spirit of Tukey's twicing, we could approximate $\boldsymbol{A}$ in the RHS of (5.18) by $\hat{\boldsymbol{A}}_{\mathrm{LS}}$, which leads to a new estimator

$$\hat{\boldsymbol{A}}_{\mathrm{T}}^{(1)} = \hat{\boldsymbol{A}}_{\mathrm{LS}} + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}).$$

In fact, there exists a family of estimators by approximating $\boldsymbol{A}$ recursively in the RHS of (5.18) by $\hat{\boldsymbol{A}}_{\mathrm{T}}^{(t-1)}$ (with $\hat{\boldsymbol{A}}_{\mathrm{T}}^{(0)} = \hat{\boldsymbol{A}}_{\mathrm{LS}}$):

$$\hat{\boldsymbol{A}}_{\mathrm{T}}^{(t)} = \hat{\boldsymbol{A}}_{\mathrm{LS}} + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*(\hat{\boldsymbol{A}}_{\mathrm{T}}^{(t-1)} - \boldsymbol{B}). \tag{5.19}$$

This family of estimators can be explicitly written as

$$\hat{\boldsymbol{A}}_{\mathrm{T}}^{(t)} = \boldsymbol{B} + \boldsymbol{U}(\mathbf{I} + \boldsymbol{T} + \boldsymbol{T}^2 + \cdots + \boldsymbol{T}^t)\boldsymbol{U}^*(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}). \tag{5.20}$$

Using $\boldsymbol{W} = (\mathbf{I} - \boldsymbol{T})^{-1} = \sum_{i=0}^{\infty} \boldsymbol{T}^i$, we have that $\boldsymbol{A}_{\mathrm{T}}^{(t)} \to \hat{\boldsymbol{A}}_{\mathrm{AT}}$ as $t \to \infty$.

In general, this family of estimators has smaller variance than $\hat{\boldsymbol{A}}_{\mathrm{AT}}$, but larger bias since they are not unbiased (see Fig. 5.2).

5.4.2 Generalization to the setting $N \neq D$

If $N > D$, then the column space of $\boldsymbol{A}$ is the same as the column space of $\boldsymbol{C}$. Let $\boldsymbol{P}$ be the projector of size $N \times D$ to this column space, then we have $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{P}^*\boldsymbol{A}$. As a result, to find an unbiased estimator of $\boldsymbol{A}$, it is sufficient to find an unbiased estimator of $\boldsymbol{P}^*\boldsymbol{A}$, which is a square matrix. Since $\boldsymbol{P}^*\boldsymbol{A}$ is close to $\boldsymbol{P}^*\boldsymbol{B}$ and $(\boldsymbol{P}^*\boldsymbol{A})(\boldsymbol{P}^*\boldsymbol{A})^* = \boldsymbol{P}^*\boldsymbol{C}\boldsymbol{P}^*$ is known, Theorem 5.4.1 is applicable, and an unbiased estimator of $\boldsymbol{P}^*\boldsymbol{A}$ can be obtained through (5.17), with $\boldsymbol{B}$ replaced by $\boldsymbol{P}^*\boldsymbol{B}$ and $\boldsymbol{C}$ replaced by $\boldsymbol{P}^*\boldsymbol{C}\boldsymbol{P}^*$. In summary, an unbiased estimator of $\boldsymbol{A}$ can be obtained in two steps:

1. Find $\hat{\boldsymbol{A}}^{(0)}_{\mathrm{AT}}$, an unbiased estimator of $\boldsymbol{P}^*\boldsymbol{A}$, by applying (5.17), with $\boldsymbol{B}$ replaced by $\boldsymbol{P}^*\boldsymbol{B}$ and $\boldsymbol{C}$ replaced by $\boldsymbol{P}^*\boldsymbol{C}\boldsymbol{P}^*$.

2. An unbiased estimator of $\boldsymbol{A}$ is obtained by $\hat{\boldsymbol{A}}_{\mathrm{AT}} = \boldsymbol{P}\hat{\boldsymbol{A}}^{(0)}_{\mathrm{AT}}$.

If $N < D$, we use the following heuristic estimator. Let $\boldsymbol{P}$ be a matrix of size $D \times N$ that is the projector to the row space of $\boldsymbol{B}$, and assuming that $\hat{\boldsymbol{A}}$, the estimator of $\boldsymbol{A}$, has the same row space as $\boldsymbol{B}$, then $\hat{\boldsymbol{A}} = \hat{\boldsymbol{A}}\boldsymbol{P}\boldsymbol{P}^*$, and it is sufficient to find $\hat{\boldsymbol{A}}\boldsymbol{P}$, an estimator of $\boldsymbol{A}\boldsymbol{P}$. With $(\boldsymbol{A}\boldsymbol{P})(\boldsymbol{A}\boldsymbol{P})^* = \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{C}$ known and the fact that $\boldsymbol{A}\boldsymbol{P}$ is close to $\boldsymbol{B}\boldsymbol{P}$, we may use the estimator (5.17). In summary, we use the following procedure:

1. Find $\hat{\boldsymbol{A}}^{(0)}_{\mathrm{AT}}$, an estimator of $\boldsymbol{A}\boldsymbol{P}$, by applying the estimator (5.17), with $\boldsymbol{B}$ replaced by $\boldsymbol{B}\boldsymbol{P}$.

2. An estimator of $\boldsymbol{A}$ is obtained by $\hat{\boldsymbol{A}}_{\mathrm{AT}} = \hat{\boldsymbol{A}}^{(0)}_{\mathrm{AT}}\boldsymbol{P}^*$.

We remark that for $N < D$ there is no theoretical guarantee to show that it is an unbiased estimator, unlike the setting $N \geq D$. However, the assumption that $\hat{\boldsymbol{A}}$ has the same column space as $\boldsymbol{B}$ is reasonable, and the proposed estimator performs well in practice.

5.5 Proof of Theorem 5.4.1

5.5.1 Explicit expression of $\hat{\boldsymbol{A}}_{\mathbf{LS}}$

Since $\hat{\boldsymbol{A}}_{\mathrm{LS}}$ is independent of the choice of $\boldsymbol{F}$ in the algorithm, we may assume that $\boldsymbol{F} = \boldsymbol{A}$ without loss of generality. Let $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{B}$, then by assumption, $\boldsymbol{E}$ is fixed, and

$$\begin{aligned}\boldsymbol{V}_0\boldsymbol{U}_0^* =&(\boldsymbol{V}_0\boldsymbol{\Sigma}_0\boldsymbol{U}_0^*)(\boldsymbol{U}_0\boldsymbol{\Sigma}_0^{-1}\boldsymbol{U}_0^*) = (\boldsymbol{V}_0\boldsymbol{\Sigma}_0\boldsymbol{U}_0^*)[\boldsymbol{U}_0\boldsymbol{\Sigma}_0^2\boldsymbol{U}_0^*]^{-0.5}\\ =&\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})[(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})]^{-0.5}.\end{aligned}$$

Therefore,

$$\hat{\boldsymbol{A}}_{\mathrm{LS}} = \boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})[(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})]^{-0.5}. \tag{5.21}$$

Applying (5.21), we may simplify $\hat{\boldsymbol{A}}_{\mathrm{LS}}$ further as follows:

$$\begin{aligned}\hat{\boldsymbol{A}}_{\mathrm{LS}} =&(\boldsymbol{A}-\boldsymbol{E})^{*-1}(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})[(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})]^{-0.5}\\ =&(\boldsymbol{A}-\boldsymbol{E})^{*-1}[(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})]^{0.5}.\end{aligned} \tag{5.22}$$

Since $\boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{C}$, we may assume that the SVD decomposition of $\boldsymbol{A}$ be given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_D)$. Let $\boldsymbol{E}_0 = \boldsymbol{U}^*\boldsymbol{E}\boldsymbol{V}$, then applying the derivative of matrix inversion we have

$$\begin{aligned}(\boldsymbol{A}-\boldsymbol{E})^{*-1} =&\boldsymbol{A}^{*-1} + \boldsymbol{A}^{*-1}\boldsymbol{E}^*\boldsymbol{A}^{*-1} + O(\|\boldsymbol{E}\|_F^2)\\ =&\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^* + (\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*)\boldsymbol{E}^*(\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*) + O(\|\boldsymbol{E}\|_F^2)\\ =&\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^* + \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^* + O(\|\boldsymbol{E}\|_F^2).\end{aligned} \tag{5.23}$$

We also have

$$(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}=\boldsymbol{A}^*\boldsymbol{A}-\boldsymbol{E}^*\boldsymbol{A}=\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^*-\boldsymbol{E}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*=\boldsymbol{V}[\boldsymbol{\Sigma}^2-\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^*$$

and similarly, $\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})=\{\boldsymbol{V}[\boldsymbol{\Sigma}^2-\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^*\}^*=\boldsymbol{V}[\boldsymbol{\Sigma}^2-\boldsymbol{\Sigma}\boldsymbol{E}_0]\boldsymbol{V}^*$. Then

$$\begin{aligned}\{(\boldsymbol{A}-\boldsymbol{E})^*\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{A}-\boldsymbol{E})\}^{0.5}=&\{\boldsymbol{V}[\boldsymbol{\Sigma}^2-\boldsymbol{E}_0^*\boldsymbol{\Sigma}][\boldsymbol{\Sigma}^2-\boldsymbol{\Sigma}\boldsymbol{E}_0]\boldsymbol{V}^*\}^{0.5}\\=&\boldsymbol{V}\{[\boldsymbol{\Sigma}^2-\boldsymbol{E}_0^*\boldsymbol{\Sigma}][\boldsymbol{\Sigma}^2-\boldsymbol{\Sigma}\boldsymbol{E}_0]\}^{0.5}\boldsymbol{V}^*\\=&\boldsymbol{V}[\boldsymbol{\Sigma}^4-\boldsymbol{E}_0^*\boldsymbol{\Sigma}^3-\boldsymbol{\Sigma}^3\boldsymbol{E}_0+O(\|\boldsymbol{E}\|_F^2)]^{0.5}\boldsymbol{V}^*\end{aligned}\tag{5.24}$$

Applying Lemma 5.5.1, we have that

$$[\boldsymbol{\Sigma}^4-\boldsymbol{E}_0^*\boldsymbol{\Sigma}^3-\boldsymbol{\Sigma}^3\boldsymbol{E}_0+o(\|\boldsymbol{E}\|_F)]^{0.5}=\boldsymbol{\Sigma}^2+\boldsymbol{Z}+o(\|\boldsymbol{E}\|_F),\tag{5.25}$$

where the ij-th entry of $\boldsymbol{Z}$ is given by

$$\boldsymbol{Z}_{ij}=-\frac{\boldsymbol{E}_{0,ji}^*\sigma_j^3+\boldsymbol{E}_{0,ij}\sigma_i^3}{\sigma_i^2+\sigma_j^2}.$$

Combining (5.22)-(5.25), we have

$$\begin{aligned}\hat{\boldsymbol{A}}_{\mathrm{LS}}=&[\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*+\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*]\boldsymbol{V}[\boldsymbol{\Sigma}^2+\boldsymbol{Z}]\boldsymbol{V}^*+o(\|\boldsymbol{E}\|_F)\\=&\boldsymbol{A}+[\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*]\boldsymbol{V}\boldsymbol{Z}\boldsymbol{V}^*+\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^*\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^*+o(\|\boldsymbol{E}\|_F)\\=&\boldsymbol{A}+\boldsymbol{U}[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^*+o(\|\boldsymbol{E}\|_F),\end{aligned}\tag{5.26}$$

and the ij-th entry of $[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}+\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]$ can be explicitly written down by

$$\begin{aligned}\sigma_i^{-1}\boldsymbol{Z}_{ij}+\sigma_i^{-1}\boldsymbol{E}_{0,ji}^*\sigma_j=&-\frac{\boldsymbol{E}_{0,ji}^*\sigma_j^3+\boldsymbol{E}_{0,ij}\sigma_i^3}{\sigma_i(\sigma_i^2+\sigma_j^2)}+\boldsymbol{E}_{0,ji}^*\frac{\sigma_j}{\sigma_i}\\=&\boldsymbol{E}_{0,ji}^*\frac{\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2}-\boldsymbol{E}_{0,ij}\frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}.\end{aligned}$$

5.5.2 Expectation when $\boldsymbol{V}$ is uniformly distributed

From the analysis in the previous section, we have

$$\mathbb{E}(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{A}) = \boldsymbol{U}\ \mathbb{E}_{\boldsymbol{V}}[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^* + o(\|\boldsymbol{E}\|_F),$$

when $\boldsymbol{V}$ is uniformly distributed on the set of all orthogonal matrices or all unitary matrices.

Now let us check the ik-th entry of $\mathbb{E}_{\boldsymbol{V}}[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^*$, which is

$$\begin{aligned}
&\sum_j \Big[\boldsymbol{E}^*_{0,ji}\frac{\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2} - \boldsymbol{E}_{0,ij}\frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}\Big]\boldsymbol{V}^*_{kj} \\
=&\sum_j \Big[(\sum_{m,n}\boldsymbol{U}^*_{mj}\boldsymbol{E}_{mn}\boldsymbol{V}_{ni})^*\frac{\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2} - \sum_{m,n}\boldsymbol{U}^*_{mi}\boldsymbol{E}_{mn}\boldsymbol{V}_{nj}\frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}\Big]\boldsymbol{V}^*_{kj}.
\end{aligned}$$

Applying the facts that for real-values matrices $\boldsymbol{V}$ we have

$$\mathbb{E}_{\boldsymbol{V}}\,\boldsymbol{V}_{ij}\boldsymbol{V}_{mn} = \begin{cases} \frac{1}{D},\ \text{if } i=m \text{ and } j=n \\ 0,\ \text{otherwise,} \end{cases}$$

and for complex-valued matrices $\boldsymbol{V}$, $\mathbb{E}_{\boldsymbol{V}}\,\boldsymbol{V}_{ij}\boldsymbol{V}_{mn} = 0$ for all (i,j,m,n), and

$$\mathbb{E}_{\boldsymbol{V}}\,\boldsymbol{V}_{ij}\boldsymbol{V}^*_{mn} = \begin{cases} \frac{1}{D},\ \text{if } i=m \text{ and } j=n \\ 0,\ \text{otherwise,} \end{cases}$$

its expectation is given by

$$\frac{1}{D}\Big\{\sum_m (\boldsymbol{U}^*_{mi}\boldsymbol{E}_{mk})^*\frac{1}{2} - \sum_{m,j}\boldsymbol{U}^*_{mi}\boldsymbol{E}_{mk}\frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}\Big\}$$
$$= \begin{cases} \frac{1}{D}\Big\{\frac{1}{2}[\boldsymbol{U}^*\boldsymbol{E}]_{ik} - [\boldsymbol{U}^*\boldsymbol{E}]_{ik}\sum_j \frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}\Big\}, & \text{when } \boldsymbol{V} \text{ is real-valued,} \\ -\frac{1}{D}[\boldsymbol{U}^*\boldsymbol{E}]_{ik}\sum_j \frac{\sigma_i^2}{\sigma_i^2+\sigma_j^2}, & \text{when } \boldsymbol{V} \text{ is complex-valued.} \end{cases}$$

Combining these elementwise expectations into a matrix, $\mathbb{E}_{\boldsymbol{V}}[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} - \boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^* = -\boldsymbol{T}\boldsymbol{U}^*\boldsymbol{E}$. Therefore, we have

$$\mathbb{E}(\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{A}) = \boldsymbol{U}\ \mathbb{E}_{\boldsymbol{V}}[\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} - \boldsymbol{\Sigma}^{-1}\boldsymbol{E}_0^*\boldsymbol{\Sigma}]\boldsymbol{V}^* = -\boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*\boldsymbol{E} + o(\|\boldsymbol{E}\|_F). \tag{5.27}$$

5.5.3 Lemmas

Lemma 5.5.1. *For a diagonal matrix $\boldsymbol{X} = \mathrm{diag}(x_1, x_2, \cdots, x_D)$, the ij-th entry of $(\boldsymbol{X}+\boldsymbol{E})^{0.5} - \boldsymbol{X}^{0.5}$ is given by*

$$[(\boldsymbol{X}+\boldsymbol{E})^{0.5} - \boldsymbol{X}^{0.5}]_{ij} = [\boldsymbol{E}_{ij}\cdot\frac{1}{x_i^{0.5}+x_j^{0.5}}] + o(\|\boldsymbol{E}\|_F).$$

Proof. The proof is based on the following observation: if $\boldsymbol{Y}$ is diagonal and $\boldsymbol{C}$ is small,

$$(\boldsymbol{Y}+\boldsymbol{C})^2 - \boldsymbol{Y}^2 = \boldsymbol{Y}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{Y} + \boldsymbol{C}^2 = [\boldsymbol{C}_{ij}\cdot(y_i+y_j)] + o(\|\boldsymbol{C}\|_F),$$

where $[\boldsymbol{C}_{ij}\cdot(y_i+y_j)]$ denotes a matrix of $D\times D$, with ij-th entry given by $\boldsymbol{C}_{ij}\cdot(y_i+y_j)$.

Then the lemma is proved by applying this observation to $\boldsymbol{Y} = \boldsymbol{X}^{0.5}$ and $(\boldsymbol{Y}+\boldsymbol{C})^2 = \boldsymbol{X}+\boldsymbol{E}$:

$$\boldsymbol{E} = \big[[(\boldsymbol{X}+\boldsymbol{E})^{0.5} - \boldsymbol{X}^{0.5}]_{ij}\cdot(x_i^{0.5}+x_j^{0.5})\big] + o(\|(\boldsymbol{X}+\boldsymbol{E})^{0.5} - \boldsymbol{X}^{0.5}\|_F).$$

□

5.6 Estimation of the Covariance and Autocorrelation Matrices

The autocorrelation matrices $\boldsymbol{C_l}$ in Kam's theory are derived from the covariance matrix $\boldsymbol{\Sigma}$ of the 2D Fourier transformed projection images through [35]

$$\boldsymbol{C_l}(|k_1|, |k_2|) = 2\pi(2l+1)\int_0^\pi \Sigma(|k_1|, |k_2|, \psi)P_l(\cos\psi)\sin\psi d\psi \tag{5.28}$$

where ψ is the angle between the vectors k_1 and k_2 in the x-y plane. We estimate the covariance matrix $\boldsymbol{\Sigma}$ of the underlying 2D Fourier transformed clean projection images using the method described in [10]. This estimation method provides a more accurate covariance compared to the classical sample covariance matrix [101, 100]. First, it corrects for the CTF. Second, it performs eigenvalue shrinkage, which is critical for high dimensional statistical estimation problems. Third, it exploits the block diagonal structure of the covariance matrix in a steerable basis, a property that follows from the fact that any experimental image is just as likely to appear in different in-plane rotations. A steerable basis consists of outer products of radial functions (such as Bessel functions) and Fourier angular modes. Each block along the diagonal corresponds to a different angular frequency [114]. Moreover, the special block diagonal structure facilitates fast computation of the covariance matrix [113].

Since the autocorrelation matrix $\boldsymbol{C_l}$ estimated from projection images can have a rank exceeding $2l+1$, we first find its best rank $2l+1$ approximation via singular value decomposition, before computing its Cholesky decomposition. In the case of symmetric molecules, we use the appropriate rank as dictated by classical representation theory of $SO(3)$ [40, 14] (less than $2l+1$).

5.6.1 Algorithm 2: Orthogonal Extension by Anisotropic Twicing

Algorithm 2 Orthogonal Extension

1: **procedure** ORTHOGONAL EXTENSION BY ANISOTROPIC TWICING (OE-AT): ESTIMATE $\boldsymbol{A}$ GIVEN $\boldsymbol{B} \approx \boldsymbol{A}$, SUBJECT TO $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^*$

2:

Input: $\boldsymbol{B} \in \mathbb{C}^{N\times D}$, $\boldsymbol{C} \in \mathbb{C}^{N\times N}$

3: Find any $\boldsymbol{F} \in \mathbb{C}^{N\times D}$ such that $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^*$

4: Calculate $\boldsymbol{B}^*\boldsymbol{F}$ and calculate its singular value decomposition $\boldsymbol{B}^*\boldsymbol{F} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$.

5: Calculate the OE-LS estimator is $\hat{\boldsymbol{A}}_{\text{LS}} = \boldsymbol{F}\boldsymbol{V}_0\boldsymbol{U}_0^*$, (see Algorithm 1).

6: For $N = D$, the OE-AT estimator is given by $\hat{\boldsymbol{A}}_{\text{AT}} = \boldsymbol{B} + \boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^*(\hat{\boldsymbol{A}_{\text{LS}}} - \boldsymbol{B})$.

7: For $N > D$, assuming that $\boldsymbol{P}$ is the projector of size $N \times D$ to the D-dimensional subspace spanned by the columns of $\boldsymbol{C}$, $\hat{\boldsymbol{A}}_{\text{AT}} = \boldsymbol{P}\hat{\boldsymbol{A}}^{(0)}_{\text{AT}}$.

8: For $N < D$, assuming that $\boldsymbol{P}$ is the projector of size $D \times N$ to the N-dimensional subspace in $\mathbb{R}^D$ spanned by the rows of $\boldsymbol{B}$, $\hat{\boldsymbol{A}}_{\text{AT}} = \hat{\boldsymbol{A}}^{(0)}_{\text{AT}}\boldsymbol{P}^*$.

5.7 Numerical Experiments

5.7.1 Bias Variance Trade-off

For any parameter θ, the performance of its estimator $\hat{\theta}$ can be measured in terms of its mean squared error (MSE), $\mathbb{E}||\theta - \hat{\theta}||^2$. The MSE of any estimator can be decomposed into its bias and variance:

$$\text{MSE} = \mathbb{E}[||\theta - \hat{\theta}||^2] = ||\text{Bias}||^2 + \text{Var} \tag{5.29}$$

where

$$\text{Bias} = \mathbb{E}[\hat{\theta}] - \theta \tag{5.30}$$

and

$$\text{Var} = \mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||^2] \tag{5.31}$$

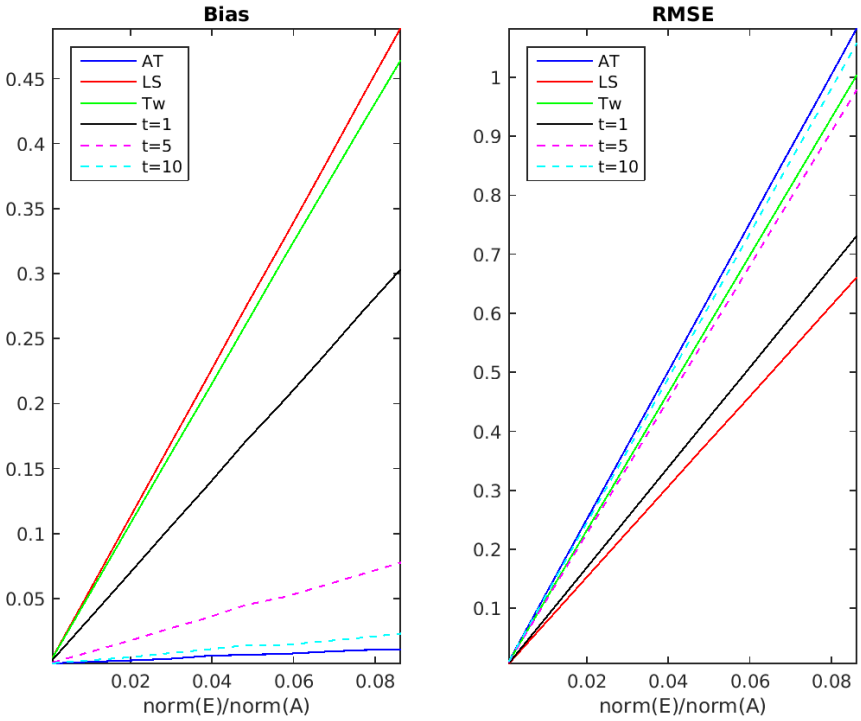


Figure 5.2: Bias and RMSE of the Anisotropic Twicing (AT), Least Squares (LS), Twicing (Tw) estimators and also the family of estimators with $t = 1, 5, 10$ averaged over 10000 experiments, as described in Sec. 5.7.1. The x-axis shows the relative perturbation $||\boldsymbol{E}||/||\boldsymbol{A}||$.

Unbiased estimators are often not optimal in terms of MSE, but they can be valuable for being unbiased. We performed a numerical experiment starting with a fixed $\boldsymbol{F} \in \mathbb{R}^{10\times 10}$ and an unknown matrix $\boldsymbol{A} = \boldsymbol{F}\boldsymbol{O}$ where $\boldsymbol{O}$ is a random orthogonal matrix. We are given a known similar matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{E}$. The goal is estimate $\boldsymbol{A}$ given $\boldsymbol{B}$ and $\boldsymbol{F}$. Figure 5.2 shows a comparison of the bias and root mean squared error (RMSE) of different estimators averaged over 10000 runs of the numerical experiment: the anisotropic twicing estimator, the twicing estimator, the least squares estimator, and estimators from the family of estimators for some values of t in Sec. 5.4.1. The figure demonstrates that the AT estimator is asymptotically unbiased at the cost of higher MSE.

5.7.2 Synthetic Dataset: Toy Molecule

We perform numerical experiments with a synthetic dataset generated from an artificial 'Mickey Mouse' molecule. The molecule $\boldsymbol{B}$ is made up of ellipsoids, and the density is set to 1 inside the ellipsoids and to 0 outside. Fig. 5.3 shows the artificial new volume $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{E}$ created by adding a small ellipsoid $\boldsymbol{E}$, which we will refer to here as the "nose", to the original mickey mouse volume $\boldsymbol{B}$. This represents the small perturbation $\boldsymbol{E}$. When the Fourier volume $\boldsymbol{B} + \boldsymbol{E}$ is expanded in the truncated spherical Bessel basis described in Sec. 5.6, the average relative perturbation $||\boldsymbol{E_l}||/||\boldsymbol{A_l}||$ for the first few coefficients in the truncated spherical harmonic expansion for $l = 1, \ldots, 10$ is 8%.

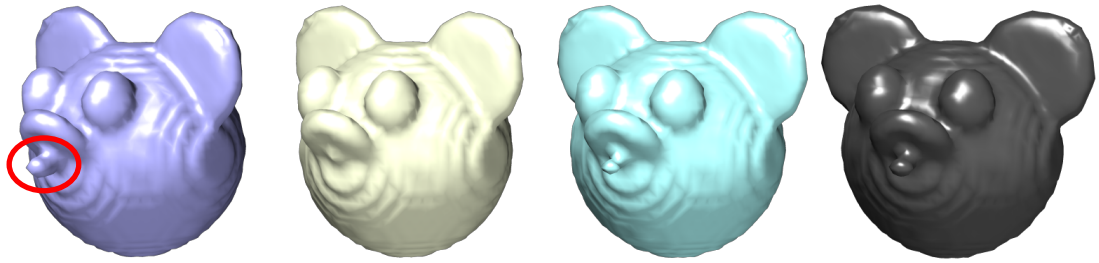
Next, we generate 10000 projection images from the volume $\boldsymbol{A}$. We then employ OE to reconstruct the volume $\boldsymbol{A}$ from $\boldsymbol{B}$ and the clean projection images of $\boldsymbol{A}$. Fig. 5.3 shows the reconstructions obtained using each of the three estimators in the OE framework, visualized in Chimera [69]. We note that while all three estimators are able to recover the additional subunit $\boldsymbol{E}$, the AT estimator best recovers the unknown subunit to its correct relative magnitude. The relative error in the region of the unknown subunit $\boldsymbol{E}$ is 59% with least squares, 31% with twicing and 19% with anisotropic twicing.

5.7.3 Synthetic Dataset: TRPV1

We perform numerical experiments with a synthetic dataset generated from the TRPV1 molecule (with imposed C_4 rotational symmetry) in complex with DkTx and RTX ($\boldsymbol{A}$). This volume is available on EMDB as EMDB-8117. The small additional subunit is visible as an extension over the top of the molecule, shown in Fig. 5.4(i). This represents the small perturbation $\boldsymbol{E}$.

Next, we generate 26000 projection images from $\boldsymbol{A}$, add the effect of both the CTF (the images are divided into 10 defocus groups) and additive white Gaussian

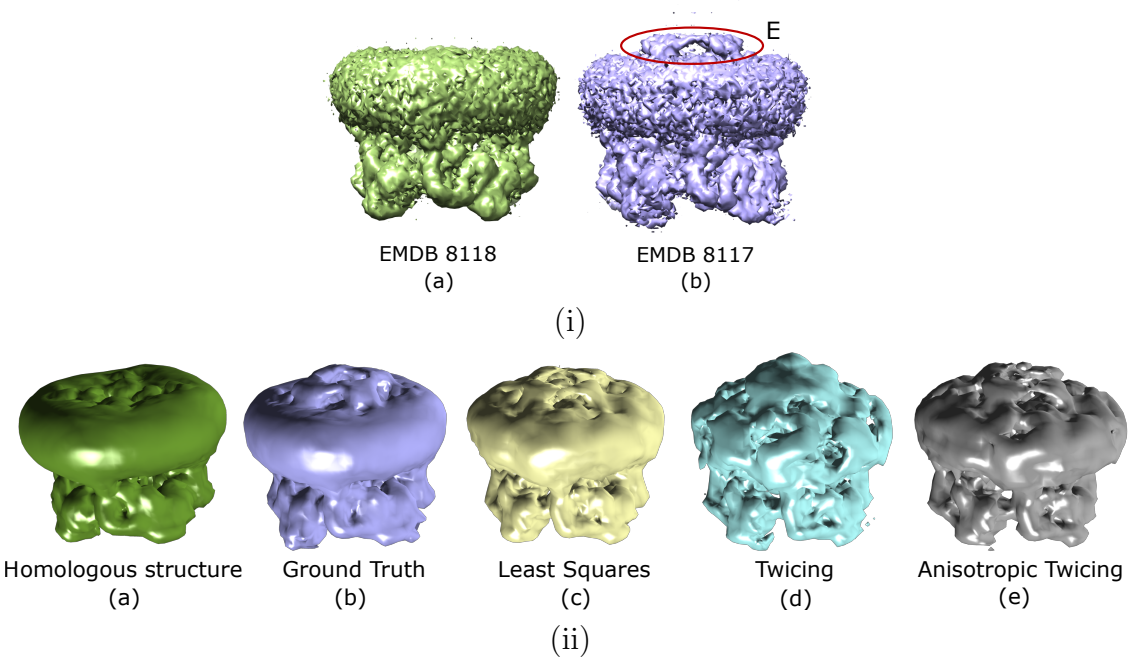
Figure 5.3: A synthetic toy mickey mouse molecule with a small additional subunit, marked 'E' in (a). We reconstruct the molecule $\boldsymbol{A}$ from its clean projection images, given $\boldsymbol{B}$. We show reconstructions obtained with the least squares estimator in (b), twicing estimator in (c), and AT estimator in (d).



Original molecule with additional subunit marked in red (a) | Least Squares (b) | Twicing (c) | Anisotropic Twicing (d)

noise (SNR=1/40) and use OE to reconstruct the volume $\boldsymbol{A}$. Fig. 5.4(ii) shows the reconstructions obtained using each of the three estimators in the OE framework, visualized in Chimera. The C_4 symmetry was taken into account in the autocorrelation analysis by including in (2) only symmetry-invariant spherical harmonics Y_l^m for which $m = 0 \mod 4$. As seen earlier with the synthetic case, all three estimators are able to recover the additional subunit $\boldsymbol{E}$, while the AT estimator best recovers the unknown subunit to its correct relative magnitude. The relative error in the unknown subunit is 43% with least squares, 56% with twicing and 30% with anisotropic twicing. We note that this is the first successful attempt, even with synthetic data, at using OE for 3D homology modeling directly from CTF-affected and noisy images (at experimentally relevant conditions). The numerical experiments using the Kv1.2 potassium channel in [9] were at an unrealistically high SNR and did not include the effect of the CTF. The reason for this improvement is the improved covariance estimation [10].

Figure 5.4: A synthetic TRPV1 molecule (EMDB 8118), with a small additional subunit DxTx and RTX (EMDB 8117), marked 'E' in (i-b). We reconstruct the molecule from its noisy, CTF-affected images, and the homologous structure. In (ii), we show reconstructions obtained with the least squares, twicing and AT estimators using OE, along with the homologous structure and the ground truth projected on to the basis in (ii-a) and (ii-b).



5.7.4 Experimental Dataset: TRPV1

We apply OE to an experimental data of the TRPV1 molecule in complex with DkTx and RTX, determined in lipid nanodisc, available on the public database Electron Microscopy Pilot Image Archive (EMPIAR) as EMPIAR-10059, and the 3D reconstruction is available on the electron microscopy data bank (EMDB) as EMDB-8117, courtesy of Y. Gao et al [24]. The dataset provided consists of 73000 motion corrected, picked particle images (which were used for the reconstruction in EMDB-8117) of size 192×192 with a pixel size $1.3\mathring{A}$. We use the 3D structure of TRPV1 alone as the similar molecule. This is available on the EMDB as EMDB-8118. The two structures differ only by the small DkTx and RTX subunit at the top, which can be seen in Fig. 5.4(i).

Since the noise in experimental images is colored while our covariance estimation procedure requires white noise, we first preprocess the raw images in order to "whiten" the noise. We estimate the power spectrum of noise using the corner pixels of all images. The images are then whitened using the estimated noise power spectrum.

In the context of our mathematical model, the volume EMDB-8117 of TRPV1 with DkTx and RTX is the unknown volume $\boldsymbol{A}$, and the volume EMDB-8118 of TRPV1 alone is the known, similar volume $\boldsymbol{B}$. We use OE to estimate $\boldsymbol{A}$ given $\boldsymbol{B}$ and the raw, noisy projection images of $\boldsymbol{A}$ from an experimental dataset.

The basis assumption in Kam's theory is that the distribution of viewing angles is uniform. This assumption is difficult to satisfy in practice, since molecules in the sample can often have preference for certain orientations due to their shape and mass distribution. The viewing angle distribution in EMPIAR-10059 is non-uniform (see Fig. 5.6). As a robustness test of our methods, we attempt 3D reconstruction with (i) all images, such that the viewing angle distribution is non-uniform, as well as (ii) by sampling images such that the viewing angle distribution of the images is approximately uniform (as shown in Fig. 5.6). We obtained the final viewing angles

estimated after refinement from the Cheng lab at UCSF [24]. Our sampling procedure is as follows: we choose 10000 points at random from the uniform distribution on the sphere and classify each image into these 10000 bins based on the point closest to it. We discard bins that have no images, and for the remaining bins we pick a maximum of 3 points per bin. We use the selected images (slighty less less than 30000) for reconstruction with roughly uniform distributed viewing angles.

The reconstructed 3D volumes are shown in Fig. 5.5. We note that the additional subunit is recovered at the right location, and roughly to the expected size, using all three estimators. This is the first instance of reconstructing a 3D model directly from raw experimental images, without any class averaging or iterative refinement, by employing OE. The Fourier cross resolution (FCR) of the reconstruction with the 'ground truth' EMDB-8117 is shown in Fig. 5.7.

The algorithm is implemented in the UNIX environment, on a machine with 60 cores, running at 2.3 GHz, with total RAM of 1.5TB. Using 20 cores, the total time taken here for preprocessing (whitening, background normalization. etc.) the 2D images and computing the covariance matrix was 1400 seconds. Calculating the autocorrelation matrices using (5.28) involves some numerical integration (eq. 7.15 in [14]) which took 790 seconds, but for a fixed c and R (satisfied for datasets of roughly similar size and quality) these can be precomputed. Computing the basis functions and calculating the coefficient matrices $\boldsymbol{A_l}$ of the homologous structure took 30 seconds and recovering the 3D structure by applying the appropriate estimator (AT, twicing, or LS) and computing the volume from the estimated coefficients took 10 seconds.

Figure 5.5: OE with an experimental data of the TRPV1 in complex with DkTx and RTX (EMPIAR-10059) whose 3D reconstruction is available as EMDB-8117. 3D reconstructions with OE using the least squares, anisotropic twicing, and twicing estimators: (i) With (slightly less than 30000) images selected by sampling to impose approximately uniform viewing angle distribution (ii) With all 73000 images such that the viewing angle distribution is non-uniform (see Fig. 5.6).

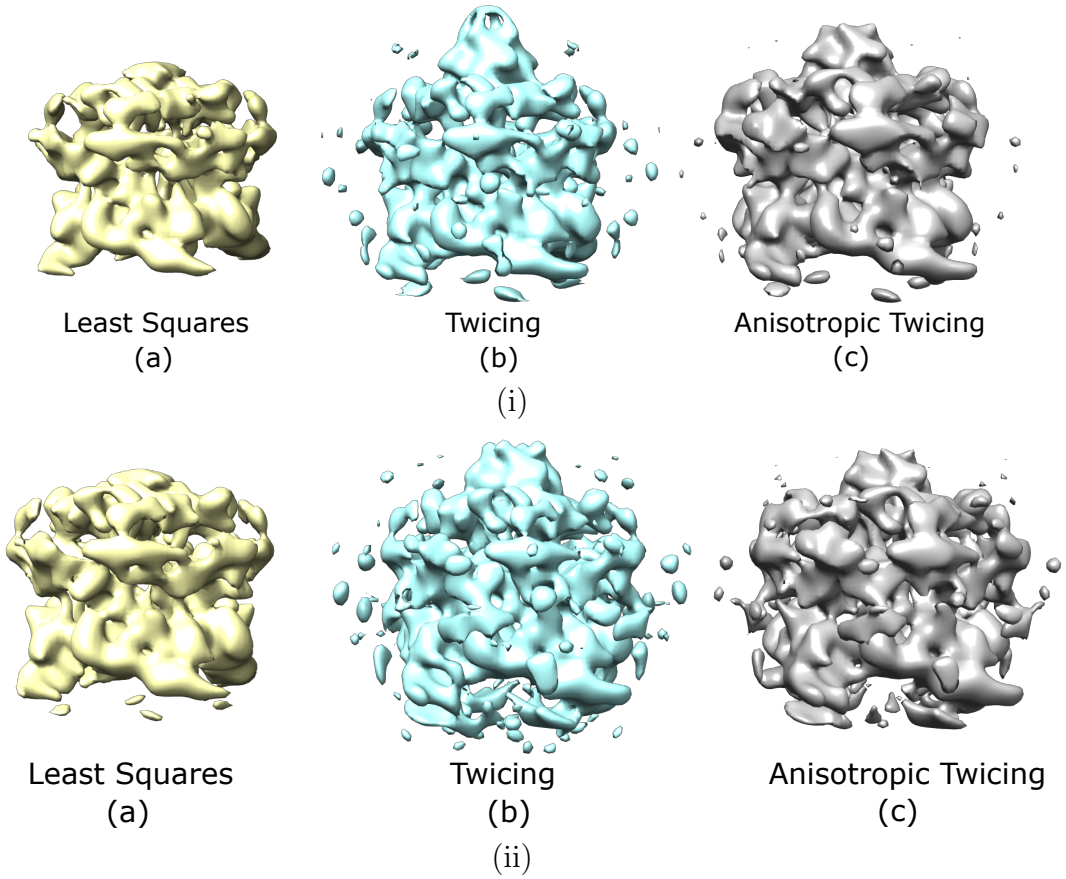
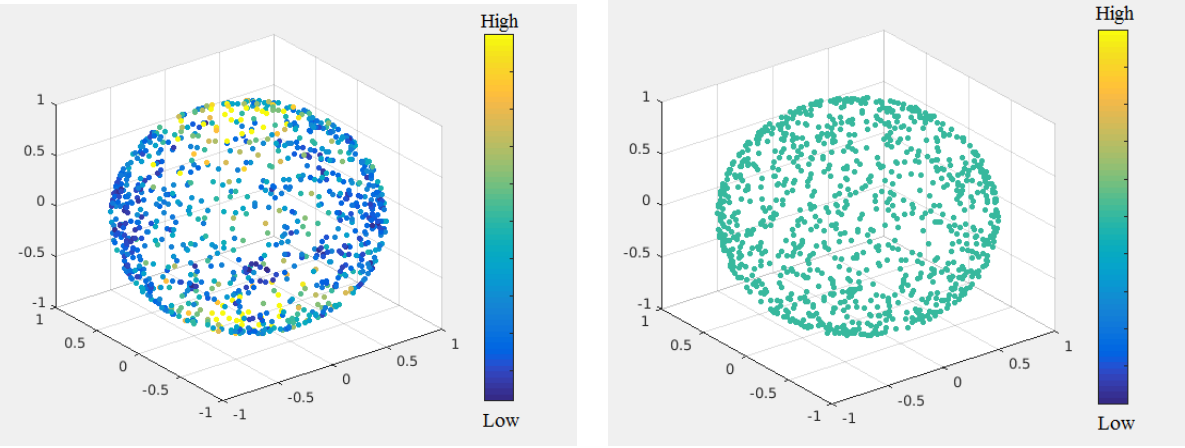
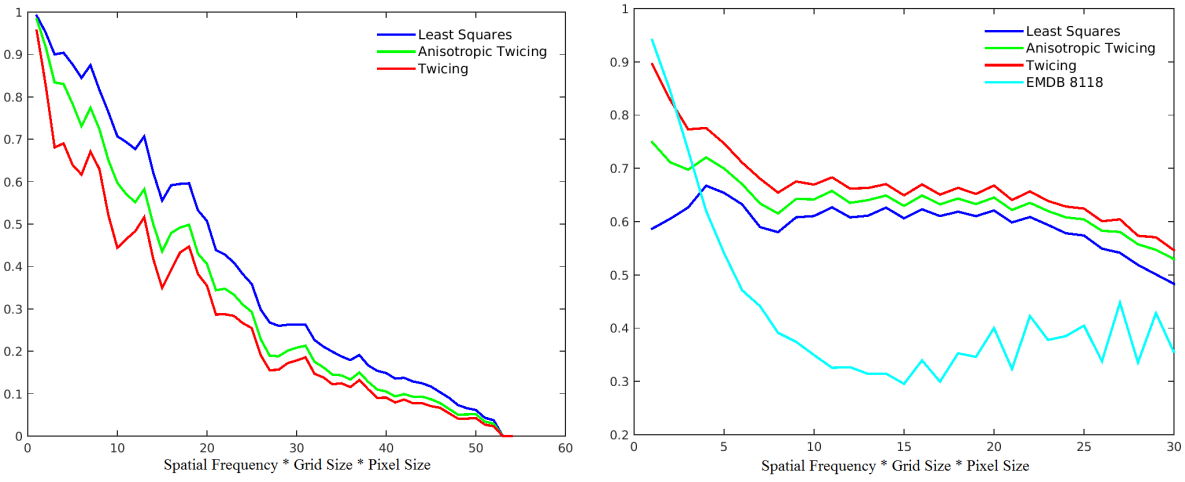


Figure 5.6: Viewing angle distribution of images in the dataset EMPIAR-10059: (i) Non-uniform distribution in the raw dataset. The visualization here shows centroids of the bins that the sphere is divided into. The color of each point is assigned based on the number of points in the bin, yellow being the largest, representing the most dense bin, and blue being the smallest. (ii) Approximately uniform distribution after sampling.



(i) (ii)

Figure 5.7: (i) FCR curve for the reconstruction of the entire molecule obtained by OE using the least squares, twicing, anisotropic twicing estimators corresponding to Fig. 5.5(i). (ii) FCR curve for the reconstruction of the unknown subunit obtained by OE using the least squares, twicing, anisotropic twicing estimators corresponding to Fig. 5.5(i). We also show the FCR of the masked homologous volume (EMDB-8118) to show the improvement in FCR obtained using OE.



(i) (ii)

5.8 Conclusion

The orthogonal retrieval problem in SPR is akin to the phase retrieval problem [12, 48, 70] in X-ray crystallography. In crystallography, the measured diffraction patterns contain information about the modulus of the 3D Fourier transform of the structure but the phase information is missing and needs to be obtained by other means. In crystallography, the particle's orientations are known but the phase of the Fourier coefficients is missing, while in cryo-EM, the projection images contain phase information but the orientations of the particles are missing. Kam's autocorrelation analysis for SPR leads to an orthogonal retrieval problem which is analogous to the phase retrieval problem in crystallography. The phase retrieval problem is perhaps more challenging than the orthogonal matrix retrieval problem in cryo-EM. In crystallography each Fourier coefficient is missing its phase, while in cryo-EM only a single orthogonal matrix is missing per several radial components. For each l, the unknown coefficient matrix $\boldsymbol{A_l}$ is of size $S_l \times (2l+1)$, corresponding to $(2l+1)$ radial functions. Each $\boldsymbol{A_l}$ is to be obtained from $\boldsymbol{C_l}$, which is a positive semidefinite matrix of size $S_l \times S_l$ and rank at most $2l+1$. For $S_l > 2l+1$, instead of estimating $S_l(2l+1)$ coefficients, we only need to estimate an orthogonal matrix in $\mathrm{O}(2l+1)$ which allows $l(2l+1)$ degrees of freedom. Therefore there are $(S_l - l)(2l+1)$ fewer parameters to be estimated.

It is important to note that the main requirement for OE to succeed is that there are sufficiently many images to estimate the covariance matrix to the desired level of accuracy, so it has a much greater chance of success for homology modeling from very noisy images than other ab-initio methods such as those based on common lines, which fail at very high noise levels.

In this paper, we find a general magnitude correction scheme for the class of 'phase-retrieval' problems, in particular, for Orthogonal Extension in cryo-EM. The magnitude correction scheme is a generalization of 'twicing' that is commonly used in

molecular replacement. We derive an asymptotically unbiased estimator and demonstrate 3D homology modeling using OE with synthetic and experimental datasets. We foresee this method as a good way to provide models to initialize refinement, directly from experimental images without performing class averaging and orientation estimation in cryo-EM and XFEL.

While Anisotropic Twicing outperforms least squares and twicing for synthetic data, the three estimation methods have similar performance for experimental data. One possible explanation is that the underlying assumption made by all estimation methods that $\boldsymbol{C_l}$ are noiseless as implied by imposing the constraint $\boldsymbol{C_l} = \boldsymbol{A_l}\boldsymbol{A_l^*}$, is violated more severely for experimental data. Specifically, the $\boldsymbol{C_l}$ matrices are derived from the 2D covariance matrix of the images, and estimation errors are the result of noise in the images, finite number of images available, non-uniformity of viewing directions, and imperfect estimation of individual image noise power spectrum, contrast transfer function, and centering. These effects are likely to be more pronounced in experimental data compared to synthetic data. As a result, the error in estimating the $\boldsymbol{C_l}$ matrices from experimental data is larger. The error in the estimated $\boldsymbol{C_l}$ can be taken into consideration by replacing the constrained least squares problem (1) with the regularized least squares problem

$$\min_{\boldsymbol{A}} ||\boldsymbol{A}-\boldsymbol{B}||_F^2 + \lambda||\boldsymbol{C}-\boldsymbol{A}\boldsymbol{A}^*||_F^2 \tag{5.32}$$

where $\lambda > 0$ is a regularization parameter that would depend on the spherical harmonic order l. A comprehensive analysis of (5.32) and its application to experimental datasets will be the subject of future work.

Chapter 6

Conclusion

In this dissertation, we first presented two new approaches based on Kam's theory for homology and ab-initio modeling of macromolecules for SPR from cryo-EM. We required an estimator of the covariance matrix of 2D projection images prior to the effect of the CTF and noise. This would lead to better 2D image restoration and class averaging procedures. We estimated the covariance matrix in the presence of the CTF and realistic noise levels, and applied our methods to experimental datasets.

We presented a new approach for image restoration of cryo-EM images, CWF, whose main algorithmic components are covariance estimation and deconvolution using Wiener filtering. CWF performs both CTF correction, by correcting the Fourier phases and amplitudes of the images, as well as denoising, by eliminating the noise thereby improving the SNR of the resulting images. For future work, it remains to be seen whether the resulting denoised images from CWF can be directly used to estimate viewing angles, without performing classification and averaging. With the improvement in detector technology, it would be exciting to reach SNR's for raw images that allow direct 3D reconstruction from denoised images.

We introduced a new similarity measure to compare CTF-affected cryo-EM images belonging to different defocus groups. We provided a new probabilistic interpretation

for this anisotropic affinity. The affinity can also be used as a similarity measure for any manifold learning procedure [95, 86] such as diffusion maps [89, 15], with or without missing data, and extended to other imaging modalities where images are affected by different point spread functions or blurring kernels. It would be interesting to explore the extension of this metric from 2D images to 3D volumes, in problems where there exist distinct classes of 3D volumes. The heterogeneity problem in cryo-EM is one such instance.

Finally, we derived a general magnitude correction scheme for the class of 'phase-retrieval' problems, in particular, for Orthogonal Extension in cryo-EM. We derived an asymptotically unbiased estimator and demonstrate 3D homology modeling using OE with synthetic and experimental datasets. We foresee this method as a good way to provide models to initialize refinement, directly from experimental images without performing class averaging and orientation estimation. This method can also be extended to SPR using X-ray free electron lasers (XFEL).

Covariance matrix estimation is encountered in several statistical estimation problems across diverse disciplines, such as portfolio selection and risk management in finance, Kalman filters that occur in computer vision and vehicle navigation, inferring covariance matrices from sparse genomic data in bioinformatics, etc. The covariance estimation algorithm presented in this thesis can be extended to any problem with a similar statistical model for the data.

The Mahalanobis affinity we derived can also be applied to any manifold learning procedure such as diffusion maps, and extended to other imaging modalities where images are affected by different point spread functions or blurring kernels. The affinity measure can thus lead to an improvement in clustering and classification of data.

Chapter 7

Appendix

7.1 Singular Value Decomposition

The Singular Value Decomposition (SVD) of a matrix $A \in \mathbb{R}^{m \times n}$ is given by

$$A = U\Sigma V^T \tag{7.1}$$

where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ $(p = \min\{m, n\})$ is an $m \times n$ diagonal matrix.

The 'best' rank-k approximation of a matrix in terms of the operator norm is given by its SVD. $A_k = U\Sigma_k Y^T$, where $\Sigma_k = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k, 0, \ldots, 0)$ is a diagonal matrix containing the largest k singular values, is a rank-k matrix such that

$$||A - A_k|| = \min_{\text{rank}(B)=k} ||A - B||. \tag{7.2}$$

In fact, the SVD provides the best rank-k approximation for any unitarily invariant norm [31].

For a square $n \times n$ matrix A, its closest orthogonal matrix O in the Frobenius

norm sense [38] is given by $O = UV^T$, that is,

$$||A - UV^T||_F = \min_{OO^T = O^T O = I} ||A - O||_F. \tag{7.3}$$

7.2 High Dimensional PCA and Random Matrix Theory

We list a few results from random matrix theory and high dimensional PCA that have been employed in the work in this thesis.

Principal Component Analysis (PCA) is a linear dimensionality reduction methods popular in data analysis. The goal in PCA is to find an orthogonal transformation of the data after centering to a lower dimensional space with maximum variability captured. Given a data set that consists of n vectors $x_1, x_2, \cdots \in \mathbb{R}^p$, PCA returns a basis that is dependent on the dataset, whose elements are called the 'principal components' [28].

Typically, PCA in classical applications is restricted to the domain of fixed dimensionality p and $n \to \infty$. In most modern problems, however, p and n are comparable. This is what we call the 'high dimensionality' regime. Stein showed that the empirical covariance matrix can be improved by appropriate shrinkage of its eigenvalues [1]. Given the data matrix X whose columns are the data vectors $x_1, x_2, \ldots x_n$, the sample covariance matrix Σ_n is

$$\Sigma_n = \frac{1}{n} X X^T. \tag{7.4}$$

For random vectors $x_1, x_2, \ldots x_n \in \mathbb{R}^p$ which are i.i.d. samples drawn from $\mathbb{N}(0, I_{p \times p})$, the distribution of such $p \times p$ random matrices X is the Wishart distribution. When p is fixed and $n \to \infty$ as in the classical case, by the law of large

numbers, the eigenvalues of Σ_n concentrate around 1.

In the non-classical case, when p and n both grow such that $p/n \to \gamma$ and $0 < \gamma \leq 1$, the limiting spectral density of the eigenvalues converges to the Marčenko Pastur (MP) distribution [57], given by

$$MP(x) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{\gamma x} 1_{[\gamma_-, \gamma_+]}, \quad \gamma_\pm = (1 \pm \sqrt{\gamma})^2 \tag{7.5}$$

for $\gamma \leq 1$.

7.2.1 BPP Transition in the Spike Model

The spike model $\Sigma = I + \beta v v^T$, where v is a unit norm vector and $\beta > 0$, is used to analyze when Σ is an identity with a rank 1 perturbation. There is a critical value of β below which no change is expected to se been in the distribution of eigenvalues, and above which at least one of the eigenvalues pops out of the support. This is called the BPP transition [6].

Bibliography

[1] Some problems in multivariate analysis. *Technical report*, 1956.

[2] Method of the year 2015. *Nat Meth*, 13(1):1–1, Jan 2016. Editorial.

[3] X. Agirrezabala, H. Y. Liao, E. Schreiner, J. Fu, R. F. Ortiz-Meoz, K. Schulten, R. Green, and J. Frank. Structural characterization of mrna-trna translocation intermediates. *Proceedings of the National Academy of Sciences*, 109(16):6094–6099, 2012.

[4] J. Andén, E. Katsevich, and A. Singer. ovariance estimation using conjugate gradient for 3d classification in CRYO-EM. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 200–204, April 2015.

[5] X. C. Bai, G. McMullan, and S. H. Scheres. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.*, 40(1):49–57, Jan 2015.

[6] J. Baik, G. Ben Arous, and S. PPhase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 09 2005.

[7] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger inequality for the graph connection Laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1611–1630, 2013.

[8] A. Barnett, L. Greengard, A. Pataki, and M. Spivak. Rapid solution of the cryo-EM reconstruction problem by frequency marching. *ArXiv e-prints*, Oct. 2016.

[9] T. Bhamre, T. Zhang, and A. Singer. Orthogonal matrix retrieval in cryo-electron microscopy. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 1048–1052, April 2015.

[10] T. Bhamre, T. Zhang, and A. Singer. Denoising and covariance estimation of single particle cryo-em images. *Journal of Structural Biology*, 195(1):72 – 81, 2016.

[11] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[12] A. Burvall, U. Lundström, P. A. C. Takman, D. H. Larsson, and H. M. Hertz. Phase retrieval in x-ray phase-contrast imaging suitable for tomography. *Opt. Express*, 19(11):10359–10376, May 2011.

[13] X. chen Bai, I. S. Fernandez, G. McMullan, and S. H. Scheres. Ribosome structures to near-atomic resolution from thirty thousand cryo-em particles. In *eLife*, 2013.

[14] X. Cheng. Random matrices in high-dimensional data analysis. 2013.

[15] R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006.

[16] K. Cowtan. Kevin cowtan's picture book of fourier transforms, 2014. Last updated: 2014-10-23.

[17] J. Dainty and R. Shaw. *Image science: principles, analysis and evaluation of photographic-type imaging processes*. Academic Press, 1974.

[18] D. Donoho, M. Gavish, and I. M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *CoRR*, abs/1311.0851, 2014.

[19] P. Dube, P. Tavares, R. Lurz, and M. van Heel. Bacteriophage SPP1 portal protein: a DNA pump with 13-fold symmetry. *EMBO J.*, 15:1303–1309, 1993.

[20] N. A. Farrow and F. P. Ottensmeyer. A posteriori determination of relative projection directions of arbitrarily oriented macromolecules. *Journal of the Optical Society of America A*, 9:1749–1760, Oct. 1992.

[21] A. R. Faruqi, R. Henderson, M. Pryddetch, P. Allport, and A. Evans. Direct single electron detection with a CMOS detector for electron microscopy. *Nuclear Instruments and Methods in Physics Research A*, 546:170–175, July 2005.

[22] J. Frank. Electron microscopy of macromolecular assemblies. In J. Frank, editor, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, pages 12 – 53. Academic Press, Burlington, 1996.

[23] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies : Visualization of Biological Molecules in Their Native State: Visualization of Biological Molecules in Their Native State*. Oxford University Press, USA, 2006.

[24] Y. Gao, E. Cao, D. Julius, and Y. Cheng. Trpv1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature*, 534(5):347–51, 2016.

[25] A. B. Goncharov. Integral geometry and three-dimensional reconstruction of randomly oriented identical particles from their electron microphotos. *Acta Applicandae Mathematica*, 11(3):199–211, 1988.

[26] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, Mar. 2014.

[27] N. Grigorieff. Frealign: High-resolution refinement of single particle structures. *Journal of Structural Biology*, 157(1):117 – 125, 2007. Software tools for macromolecular microscopy.

[28] J. Han and M. Kamber. Data mining: Concepts and techniques. 2000.

[29] M. V. Heel. Angular reconstitution: a posteriori assignment of projection directions for 3d reconstruction. *Ultramicroscopy*, 21(2):111–123, 1987.

[30] R. Henderson. Realizing the potential of electron cryo-microscopy. *Quarterly Reviews of Biophysics*, 37:3–13, 2 2004.

[31] R. A. Horn and C. R. Johnson, editors. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986.

[32] A. Hosseinizadeh, A. Dashti, P. Schwander, R. Fung, and A. Ourmazd. Single-particle structure determination by x-ray free-electron lasers: Possibilities and challenges. In *Structural dynamics*, 2015.

[33] W. Jiang, M. L. Baker, Q. Wu, C. Bajaj, and W. Chiu. Applications of a bilateral denoising filter in biological electron microscopy. *Journal of structural biology*, 144 1-2:114–22, 2003.

[34] Z. Kam. Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules*, 10(5):927–934, 1977.

[35] Z. Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15 – 39, 1980.

[36] Z. Kam and I. Gafni. Three-dimensional reconstruction of the shape of human wart virus using spatial correlations. *Ultramicroscopy*, 17(3):251–262, 1985.

[37] G. Katsevich, A. Katsevich, and A. Singer. Covariance matrix estimation for the cryo-em heterogeneity problem. *SIAM J. Imaging Sciences*, 8:126–185, 2015.

[38] J. B. Keller. Closest unitary, orthogonal and hermitian operators to a given operator. *Mathematics Magazine*, 48(4):pp. 192–197, 1975.

[39] D. Kimanius, B. O. Forsberg, S. Scheres, and E. Lindahl. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *bioRxiv*, 2016.

[40] F. Klein. *Lectures on the icosahedron and the solution of equations of the fifth degree*. London, 1914. 16, 289 p.

[41] A. Klug and R. A. Crowther. Three-dimensional image reconstruction from the viewpoint of information theory. *Nature*, 238(5365):435–440, Aug 1972.

[42] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94:19–32, 2008.

[43] W. Kühlbrandt. Cryo-em enters a new era. *eLife*, 3:e03678, Aug 2014. 25122623[pmid].

[44] W. Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[45] X. Li, P. Mooney, S. Zheng, C. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng. Electron counting and beam-induced motion correction enable near atomic resolution single particle cryoem. In *Nature methods*, 2013.

[46] X. Li, S. Q. Zheng, K. Egami, D. A. Agard, and Y. Cheng. Influence of electron dose rate on electron counting images recorded with the k2 camera. *Journal of structural biology*, 184 2:251–60, 2013.

[47] M. Liao, E. Cao, D. Julius, and Y. Cheng. Structure of the trpv1 ion channel determined by electron cryo-microscopy. *Nature*, 504(7478):107?112, December 2013.

[48] Y. J. Liu, B. Chen, E. R. Li, J. Y. Wang, A. Marcelli, S. W. Wilkins, H. Ming, Y. C. Tian, K. A. Nugent, P. P. Zhu, and Z. Y. Wu. Phase retrieval in x-ray imaging based on using structured illumination. *Phys. Rev. A*, 78:023817, Aug 2008.

[49] S. J. Ludtke, T. P. Tran, Q. T. Ngo, V. Y. Moiseenkova-Bell, W. Chiu, and I. I. Serysheva. Flexible architecture of IP_3R1 by cryo-em. *Structure*, 19(8):1192 – 1199, 2011.

[50] A. V. M. Radermacher, T. Wagenknecht and J. Frank. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50s ribosomal subunit of escherichia coli. *Journal of Microscopy*, 146(2):113–136, 1987.

[51] A. V. M. Radermacher, T. Wagenknecht and J. Frank. Three-dimensional structure of the large ribosomal subunit from Escherichia coli. *EMBO J*, 6(4):1107–14, 1987.

[52] D. J. C. MacKay. Chapter 46 - deconvolution. In *Information Theory, Inference and Learning Algorithms*, pages 550 – 551. Cambridge University Press, Cambridge, UK, 2004.

[53] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, Apr. 1936.

[54] P. Main. A theoretical comparison of the β,γ' and $2F_o$–F_c syntheses. *Acta Crystallographica Section A*, 35(5):779–785, Sep 1979.

[55] S. P. Mallick, S. Agarwal, D. J. Kriegman, S. J. Belongie, B. Carragher, and C. S. Potter. Structure and view estimation for tomographic reconstruction: A bayesian approach. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2253–2260, 2006.

[56] R. Marabini, I. M. Masegosa, M. S. Martín, S. Marco, J. Fernez, L. de la Fraga, C. Vaquerizo, and J. Carazo. Xmipp: An image processing package for electron microscopy. *Journal of Structural Biology*, 116(1):237 – 240, 1996.

[57] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, Apr. 1967.

[58] J. L. Milne, M. J. Borgnia, A. Bartesaghi, E. E. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, and S. Subramaniam. Cryo-electron microscopy–a primer for the non-microscopist. *FEBS J.*, 280(1):28–45, Jan 2013.

[59] F. Natterer. *The Mathematics of Computerized Tomography*. Classics in Applied Mathematics. SIAM: Society for Industrial and Applied Mathematics, 2001.

[60] E. Nogales. The development of cryo-em into a mainstream structural biology technique. *Nat Meth*, 13(1):24–27, Jan 2016. Historical Commentary.

[61] R. Norousi, S. Wickles, C. Leidig, T. Becker, V. J. Schmid, R. Beckmann, and A. Tresch. Automatic post-picking using mappos improves particle image detection from cryo-em micrographs. *CoRR*, abs/1212.4871, 2013.

[62] W. Park and G. S. Chirikjian. An assembly automation approach to alignment of noncircular projections in electron microscopy. *IEEE Transactions on Automation Science and Engineering*, 11(3):668 – 679, 2014.

[63] W. Park, C. R. Midgett, D. R. Madden, and G. S. Chirikjian. A stochastic kinematic model of class averaging in single-particle electron microscopy. *I. J. Robotics Res.*, 30:730–754, 2011.

[64] P. Penczek, R. Renka, and H. Schomberg. Gridding-based direct Fourier inversion of the three-dimensional ray transform. *J. Opt. Soc. Am. A*, 21:499–509, 2004.

[65] P. A. Penczek. Chapter three - resolution measures in molecular electron microscopy. In G. J., editor, *Cryo-EM, Part B: 3-D Reconstruction*, volume 482 of *Methods in Enzymology*, pages 73 – 100. Academic Press, 2010.

[66] P. A. Penczek. Chapter two - image restoration in cryo-electron microscopy. In G. J. Jensen, editor, *Cryo-EM, Part B: 3-D Reconstruction*, volume 482 of *Methods in Enzymology*, pages 35 – 72. Academic Press, 2010.

[67] P. A. Penczek, M. Radermacher, and J. Frank. Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy*, 40:33–53, 1992.

[68] P. A. Penczek, J. Zhu, and J. Frank. A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously. *Ultramicroscopy*, 63(3-4):205–218, 1996.

[69] 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. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[70] F. Pfeiffer, T. Weitkamp, O. Bunk, and C. David. Phase retrieval and differential phase-contrast imaging with low-brilliance x-ray sources. *Nat Phys*, 2(4):258–261, Apr 2006.

[71] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, advance online publication, Feb. 2017.

[72] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.

[73] M. G. Rossmann. Molecular replacement – historical background. *Acta Crystallographica Section D*, 57(10):1360–1366, Oct 2001.

[74] M. G. Rossmann and D. M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, 15(1):24–31, Jan. 1962.

[75] D. K. Saldin, V. L. Shneerson, R. Fung, and A. Ourmazd. Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations. *Journal of Physics: Condensed Matter*, 21(13):134014, 2009.

[76] D. Salzman. A method of general moments for orienting 2D projections of unknown 3D objects. *Comput. Vision Graph. Image Process.*, 50(2):129–156, May 1990.

[77] B. Sander, M. Golas, and H. Stark. Advantages of {CCD} detectors for de novo three-dimensional structure determination in single-particle electron microscopy. *Journal of Structural Biology*, 151(1):92 – 105, 2005.

[78] G. Scapin. Molecular replacement then and now. *Acta Crystallographica Section D*, 69(11):2266–2275, Nov 2013.

[79] M. Schatz and M. van Heel. Invariant classification of molecular views in electron micrographs. *Ultramicroscopy*, 32:255–264, 1990.

[80] S. H. Scheres. Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519 – 530, 2012.

[81] S. H. Scheres. Semi-automated selection of cryo-em particles in relion-1.3. In *Journal of structural biology*, 2015.

[82] T. R. Shaikh, H. Gao, W. T. Baxter, F. J. A., N. Boisset, A. Leith, and J. Frank. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols*, 3(12):1941–1974, 2008.

[83] Y. Shkolnisky and A. Singer. Viewing direction estimation in cryo-em using synchronization. *SIAM J. Imaging Sciences*, 5:1088–1110, 2012.

[84] F. Sigworth. A maximum-likelihood approach to single-particle image refinement. *Journal of Structural Biology*, 122(3):328 – 339, 1998.

[85] F. J. Sigworth. Principles of cryo-em single-particle image processing. *Microscopy*, 2015.

[86] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226 – 239, 2008.

[87] A. Singer, R. R. Coifman, F. J. Sigworth, D. W. Chester, and Y. Shkolnisky. Detecting consistent common lines in cryo-EM by voting. *Journal of Structural Biology*, 2009.

[88] A. Singer and Y. Shkolnisky. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences*, 4(2):543–572, 2011.

[89] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012.

[90] C. Sorzano, S. Jonic, R. NRamz, N. Boisset, and J. Carazo. Fast, robust, and accurate determination of transmission electron microscopy contrast transfer function. *Journal of Structural Biology*, 160(2):249 – 262, 2007.

[91] C. O. Sorzano, J. Vargas, J. Ot. Abrishami, J. M. de la Rosa-Trev S. del Riego, A. Fernez-Alderete, C. Martz-Rey, R. Marabini, and J. M. Carazo. Fast and accurate conversion of atomic models into electron density maps. *AIMS Biophysics*, 2(20150102):8–20, 2015.

[92] C. O. S. Sorzano, E. Ortiz, M. López, and J. Rodrigo. Improved bayesian image denoising based on wavelets with applications to electron microscopy. *Pattern Recognition*, 39:1205–1213, 2006.

[93] D. Starodub, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, C. Bostedt, J. D. Bozek, N. Coppola, R. B. Doak, S. W. Epp, B. Erk, L. Foucar, L. Gumprecht, C. Y. Hampton, A. Hartmann, R. Hartmann, P. Holl, S. Kassemeyer, N. Kimmel, H. Laksmono, M. Liang, N. D. Loh, L. Lomb, A. V. Martin, K. Nass, C. Reich, D. Rolles, B. Rudek, A. Rudenko, J. Schulz, R. L. Shoeman, R. G. Sierra, H. Soltau, J. Steinbrener, F. Stellato, S. Stern, G. Weidenspointner, M. Frank, J. Ullrich, L. Strüder, I. Schlichting, H. N. Chapman, J. C. H. Spence, and M. J. Bogan. Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nature Comm.*, 3:1276+, Dec. 2012.

[94] H. D. Tagare, A. Kucukelbir, F. J. Sigworth, H. Wang, and M. Rao. Directly reconstructing principal components of heterogeneous particles from cryo-em images. *Journal of structural biology*, 191 2:245–62, 2015.

[95] R. Talmon and R. R. Coifman. Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proceedings of the National Academy of*

Sciences, 110(31):12535–12540, 2013.

[96] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157:38 – 46, 2007. Software tools for macromolecular microscopy.

[97] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[98] B. K. Vainshtein and A. Goncharov. Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections. *Soviet Physics Doklady*, 31:278, Apr. 1986.

[99] B. Vainstein and A. Goncharov. Determining the spatial orientation of arbitrarily arranged particles given their projections. *Dokl. Acad. Sci. USSR*, 287(5):1131–1134, 1986. English translation: Soviet Physics Doklady, Vol. 31, p.278.

[100] M. van Heel. Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy*, 13(1):165 – 183, 1984.

[101] M. van Heel and J. Frank. Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy*, 6(2):187 – 194, 1981.

[102] M. van Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Q. Rev. Biophys.*, 33(4):307–369, Nov 2000.

[103] M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz. A new generation of the IMAGIC image processing system. *Journal of Structural Biology*, 116(1):17 – 24, 1996.

[104] F. Wang, H. Gong, G. liu, M. Li, C. Yan, T. Xia, X. Li, and J. Zeng. Deeppicker: a deep learning approach for fully automated particle picking in cryo-em. *CoRR*, abs/1605.01838, 2016.

[105] J. Wang and C. Yin. A zernike-moment-based non-local denoising filter for cryo-em images. *Science China. Life sciences*, 56 4:384–90, 2013.

[106] L. Wang and F. J. Sigworth. Cryo-EM and single particles. *Physiology (Bethesda)*, 21:13–18, 2006.

[107] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953.

[108] D.-Y. Wei and C.-C. Yin. An optimized locally adaptive non-local means denoising filter for cryo-electron microscopy data. *Journal of Structural Biology*, 172(3):211 – 218, 2010.

[109] W. Wong, X.-c. Bai, A. Brown, I. S. Fernandez, E. Hanssen, M. Condron, Y. H. Tan, J. Baum, and S. H. Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, Jun 2014. 24913268[pmid].

[110] S. Xiang, F. Nie, and C. Zhang. Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600 – 3612, 2008.

[111] T. Zhang and A. Singer. Disentangling Orthogonal Matrices. *ArXiv e-prints*, June 2015.

[112] X. Zhao, Y. Li, and Q. Zhao. Mahalanobis distance based on fuzzy clustering algorithm for image segmentation. *Digit. Signal Process.*, 43(C):8–16, Aug. 2015.

[113] Z. Zhao, Y. Shkolnisky, and A. Singer. Fast steerable principal component analysis. *IEEE Transactions on Computational Imaging*, 2(1):1–12, March 2016.

[114] Z. Zhao and A. Singer. Fourier bessel rotational invariant eigenimages. *J. Opt. Soc. Am. A*, 30(5):871–877, May 2013.

[115] Z. Zhao and A. Singer. Rotationally invariant image representation for viewing direction classification in cryo-em. *Journal of Structural Biology*, 186(1):153 – 166, 2014.