Biological Imaging (2023), 3: e2
doi:10.1017/S2633903X23000028

RESEARCH ARTICLE

Fast principal component analysis for cryo-electron microscopy images

Nicholas F. Marshall[1], Oscar Mickelin[2], Yunpeng Shi[2]* and Amit Singer[23]

[1]Department of Mathematics, Oregon State University, Corvallis, Oregon 97331, USA
[2]Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544, USA
[3]Department of Mathematics, Princeton University, Princeton, New Jersey 08544, USA
*Corresponding author. E-mail: yunpengs@princeton.edu

Received: 02 November 2022; **Revised:** 13 January 2023; **Accepted:** 25 January 2023

Keywords: Covariance estimation; cryo-EM; denoising; Fourier–Bessel; principal component analysis; single particle reconstruction

Abstract

Principal component analysis (PCA) plays an important role in the analysis of cryo-electron microscopy (cryo-EM) images for various tasks such as classification, denoising, compression, and ab initio modeling. We introduce a fast method for estimating a compressed representation of the 2-D covariance matrix of noisy cryo-EM projection images affected by radial point spread functions that enables fast PCA computation. Our method is based on a new algorithm for expanding images in the Fourier–Bessel basis (the harmonics on the disk), which provides a convenient way to handle the effect of the contrast transfer functions. For N images of size $L \times L$, our method has time complexity $O\left(NL^3 + L^4\right)$ and space complexity $O\left(NL^2 + L^3\right)$. In contrast to previous work, these complexities are independent of the number of different contrast transfer functions of the images. We demonstrate our approach on synthetic and experimental data and show acceleration by factors of up to two orders of magnitude.

Impact Statement

Singe particle reconstruction in cryo-electron microscopy (cryo-EM) is an increasingly popular technique for near-atomic imaging of biological macromolecules. Both technological and computational advances have driven the progress of the techniques, yet many computational obstacles still remain. We introduce a fast method to estimate the covariance matrix of noisy cryo-EM images, which is a central component of many computational cryo-EM techniques. As an application, we use the covariance matrix for image denoising and deconvolution of the microscope's contrast transfer function. Our method provides orders of magnitude speedup compared to previous approaches, which opens the door to tackling more challenging datasets.

1. Introduction

We study the problem of computing a compressed representation of the covariance matrix of 2-D cryo-electron microscopy (cryo-EM) images for the purpose of performing principal component analysis (PCA). More precisely, we consider an image formation model where the measurement g_i is defined by

$$g_i = h_i * f_i + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, N, \tag{1}$$

where h_i is a radial function, $*$ denotes convolution, f_i is a ground-truth image, ε_i is the noise term, and N the total number of images. We emphasize that we assume h_i is radial, see Assumption A1.

We are motivated by single particle cryo-EM imaging, which is an important technique for determining the 3-D structure of macromolecules. In particular, the single particle reconstruction (SPR) problem asks to recover the 3-D structure of a macromolecule from noisy 2-D images of its tomographic projections along unknown viewing angles. In cryo-EM, the mathematical model is a special case of (1) and is of the form

$$g_i(x') = h_i * \int_{\mathbb{R}} \varphi_i\left(R_i^{-1}x\right)dx_3 + \varepsilon_i(x'), \quad i = 1,\ldots,N, \tag{2}$$

where $x = (x', x_3) \in \mathbb{R}^2 \times \mathbb{R} \cong \mathbb{R}^3$ are 3-D spatial coordinates with x_3 representing the projection direction, h_i is the point spread function, $\varphi_i : \mathbb{R}^3 \to \mathbb{R}$ is the electrostatic potential of a molecule, $R_i \in SO(3)$ is a 3-D rotation, and ε_i the noise term. In computational microscopy, it is typical to work with the Fourier transform of the point spread function, which is known as the contrast transfer function (CTF). In the simplest case, each measurement could correspond to a single fixed molecule potential function $\varphi_i = \varphi$; however, in general, we may assume that each φ_i could be a random variable representing a mixture of molecules, conformational heterogeneity, cases where the images are not perfectly centered, or other measurement imperfections[1,2].

In general, each measurement g_i can be associated with a different point spread function; however, in practice, a group of measurements, called a defocus group, can share a common point spread function. We assume that the measurements are grouped into $M \leq N$ defocus groups. Given g_i and h_i for $i = 1,\ldots,N$, our goal is to estimate the 2-D covariance function $c : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ of the images

$$c(x', y') := \mathbb{E}\left[\left(f(x') - \overline{f}(x')\right)\left(f(y') - \overline{f}(y')\right)\right], \tag{3}$$

where f is a random variable from the same distribution as the images f_i, and $\overline{f}(x') = \mathbb{E}[f(x')]$. We assume that the distribution of the images is invariant to in-plane rotations (which is typically the case in cryo-EM).

In cryo-EM, the random variable f is of the form $f(x') = \int_{\mathbb{R}} \Phi\left(R^{-1}x\right)dx_3$, where the random variable R is an unknown viewing angle, and the random variable Φ is a molecule potential. There is generally no physical reason for a molecule to prefer one in-plane rotation to another so distributions of random variables of this form are generally invariant to in-plane rotations. In the field of cryo-EM processing, the covariance function c is simply referred to as the 2-D covariance.

1.1. Motivation

The 2D-covariance is an essential component of a number of computational techniques in cryo-EM; we survey a few of these below.

First, we are motivated by PCA, which is a ubiquitous technique in statistics, data science, and computational mathematics and has applications to dimensionality reduction, denoising, visualization, among others. The principal components (i.e., the top eigenvectors of the digitized covariance matrix) have a number of uses in the computational cryo-EM pipeline. The subspace corresponding to the top eigenvectors of the covariance matrix identifies salient features of the dataset which enables, for instance, improved methods for image classification and visualization, such as Multivariate Statistical Analysis[3–6]. These techniques improve computational speed, since clustering becomes computationally easier in a space of reduced dimension, as well as accuracy, since dimensionality reduction by PCA amplifies the effective signal-to-noise ratio (SNR) because many coordinates for which noise dominates the signal are eliminated[7].

Second, the covariance matrix has applications in the method of moments, a classical statistical inference method, applied to cryo-EM[8]. In this method, the 2-D covariance is used to compute the similarly defined autocorrelation function of the underlying 3-D structure. Under further assumptions such as sufficient nonuniformity of the distribution of the viewing angles[9] or sufficient sparsity of the molecular structure[10], this autocorrelation function determines the 3-D density map either up to a finite list of possible structures or uniquely, respectively. This has been further developed into principled

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

methods for ab initio estimation of cryo-EM structures[9,10], with reduced risk of user-induced model bias in the initial model. Alternatively, when additional information is available, for instance, one[11] or two[12] noiseless projection images, or the 3-D structure of a related, homologous structure[13,14], the 3-D density map is uniquely determined by the autocorrelation, without requiring any structural assumptions.

Third, the covariance matrix has applications in denoising and CTF-correcting projection images. Covariance Wiener Filtering (CWF)[15] is an approach that uses the classic Wiener filtering framework with the estimated covariance matrix to solve the image deconvolution and denoising problem. The technique represents images in a lower dimensional subspace that is formed from PCA using the estimated covariance matrix. The method then applies Wiener filtering to correct the CTFs and denoise the images in this reduced subspace.

Compared to the standard PCA problem, the cryo-EM setting exhibits further computational challenges, since the estimation method also has to account for convolution with the point spread function, which destroys information of the resulting convolved function; see Section Section 2.2 for a more precise statement. On the other hand, the problem has additional symmetries making fast algorithms possible. In this paper, we present a new fast algorithm for estimating the covariance matrix that improves upon past approaches (especially when there are a large number of defocus groups) in terms of time and space complexity.

1.2. Main contribution

The main contribution of this paper is a new computational method for estimating the covariance Equation (3) from N measurements of the form Equation (1) encoded by $L \times L$ digitized images. The presented fast method has time complexity $O\left(NL^3 + L^4\right)$ independent of the number $M \leq N$ of defocus groups. This is in contrast to past methods, where this complexity scales poorly with M and involves $O\left(MTL^4 + NL^3\right)$ operations[15], where T is the number of iterations needed in a Conjugate Gradient step. Many modern cryo-EM experimental datasets fall into the computationally challenging regime where M scales with N.

Our fast method hinges on a new fast and accurate method for expanding $L \times L$ images into the Fourier–Bessel basis, which provides a convenient way to handle convolution of radial functions (such as point spread functions) with images: namely, convolution with radial functions can be expanded as a diagonal operator operating on the basis coefficients[16].

The Fourier–Bessel basis functions are *harmonics* on the disk: the standing waves associated with the resonant frequencies of a disk-shaped drum with a fixed boundary. More precisely, the harmonics on the disk are eigenfunctions of the Laplacian on the unit disk that satisfy Dirichlet boundary conditions. In computational mathematics, this basis is referred to as the Fourier–Bessel basis, since the basis functions can be expressed as a product of a Bessel function and a complex exponential; see Equation (6) for a definition.

Because of this simple structure, the covariance matrix of clean images can be estimated by a simple closed-form solution, without using the (computationally expensive) conjugate gradient method from previous approaches. Simultaneously, the covariance matrix retains its block diagonal structure, meaning that its diagonal blocks can be estimated separately and independently, which altogether makes PCA fast.

We present numerical results of covariance estimation on synthetic and experimental data. Additionally, we show how the estimated covariance matrix can be used to denoise images using CWF, and perform PCA to visualize eigenimages from experimental data. Code implementing the method is publicly available online.[1] Moreover, our approach has the potential to generalize to settings beyond cryo-EM, where PCA is used for signals estimated under more general group actions[17].

The remainder of the article is organized as follows. In Section 2, we describe the computational method. In Section 3, we present numerical results for synthetic data. In Section 4 we present numerical results for experimental data. In Section 5, we discuss the results and possible extensions.

[1] Code is available at https://github.com/yunpeng-shi/fast-cryoEM-PC.A.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

2. Methodology

2.1. Notation

For two $M \times N$-matrices A and B, we denote their Hadamard (or entrywise) product by $A \odot B$, the Hadamard division of A and B by $A \oslash B$ and the ℓth Hadamard power of A by $A^{\odot \ell}$. These operations are defined elementwise by

$$(A \odot B)_{jk} = A_{jk} B_{jk}, \quad (A \oslash B)_{jk} = \frac{A_{jk}}{B_{jk}}, \quad \left(A^{\odot \ell}\right)_{jk} = A_{jk}^{\ell}, \tag{4}$$

respectively. If w is an N-dimensional vector, then $\mathrm{diag}(w)$ denotes the $N \times N$ matrix with w along its diagonal, that is, $\mathrm{diag}(w)_{jj} = w_j$, and zeros elsewhere. If $f : \mathbb{R}^2 \to \mathbb{R}$ is a radial function, we write $f(x) = f(|x|)$ to mean that f can be expressed as a function only of the magnitude $|x|$ of x.

For an integrable function $f : \mathbb{R}^2 \to \mathbb{R}$, we denote the Fourier transform of f by $\widehat{f} : \mathbb{R}^2 \to \mathbb{C}$ with the convention that

$$\widehat{f}(\xi) = \frac{1}{2\pi} \int_{\mathbb{R}^2} f(x) e^{-ix \cdot \xi} dx. \tag{5}$$

2.2. Technical details

We make the following assumptions

A1. We assume that the point spread functions h_i in the model Equation (1) are radial functions; this implies that their Fourier transform (the CTFs) are also radial. In systems where astigmatism is present and the point spread function deviates slightly from a radial function, our approach can be used as an initial approximation that could be refined using the Conjugate Gradient method.

A2. We assume that the underlying images f_i in the model Equation (1) are i.i.d. random variables whose distribution is invariant to in-plane rotations.

A3. We assume a technical condition on the Fourier transform of the point spread functions h_i in the model Equation (1). Namely, that the Fourier transforms $\widehat{h}_i$ of the h_i satisfy

$$\inf_{|\xi|, |\eta| \in [\lambda_{\min}, \lambda_{\max}]} \sum_{i=1}^{N} \left|\widehat{h}_i(\xi)\right|^2 \left|\widehat{h}_i(\eta)\right|^2 \geq \delta,$$

where $\delta > 0$ is fixed and $[\lambda_{\min}, \lambda_{\max}]$ is the interval of Fourier space used in the disk harmonic expansion, see Ref. (16, Section 2.4).

Informally speaking, Assumption A3 can be interpreted as saying that for any pair of frequencies ξ and η there is a point spread function h_i such that neither $\widehat{h}_i(\xi)$ nor $\widehat{h}_i(\eta)$ are zero. Assumption A1 implies that $\widehat{h}_i(\xi) = \widehat{h}_i(|\xi|)$ is a radial function. Figure 1 shows the values of $\sum_{i=1}^{N} \left|\widehat{h}_i(\xi)\right|^2 \left|\widehat{h}_i(\eta)\right|^2$ for each pair of radial frequency ξ, η in log scale, for 1081 distinct CTF images of size $L = 360$, whose defocus values range from 0.81 to 3.87 m, for an experimental dataset; see Section 4 for more details.

This assumption is much weaker than assuming that for all i that $|\widehat{h}_i|$ does not vanish at any frequency. In the latter case, we could just use h_i to invert each equation to get access to the underlying functions f_i. If we had direct access to the underlying functions f_i, then we could approximate the covariance function c by the sample covariance

$$c_N(x', y') := \frac{1}{N-1} \sum_{i=1}^{N} \left[f_i(x') - \tilde{f}(x')\right] \cdot \left[f_i(y') - \tilde{f}(y')\right],$$

where $\tilde{f}(x') = \sum_{i=1}^{N} \frac{1}{N} f_i(x')$ is the sample mean. Indeed, $c_N(x', y') \to c(x', y')$ by the law of large numbers. To clarify why it is useful for $\widehat{h}_i$ not to vanish, note that in the Fourier domain our measurement model can be expressed as

$$\widehat{g}_i = \widehat{h}_i \widehat{f}_i + \widehat{\varepsilon}_i,$$

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

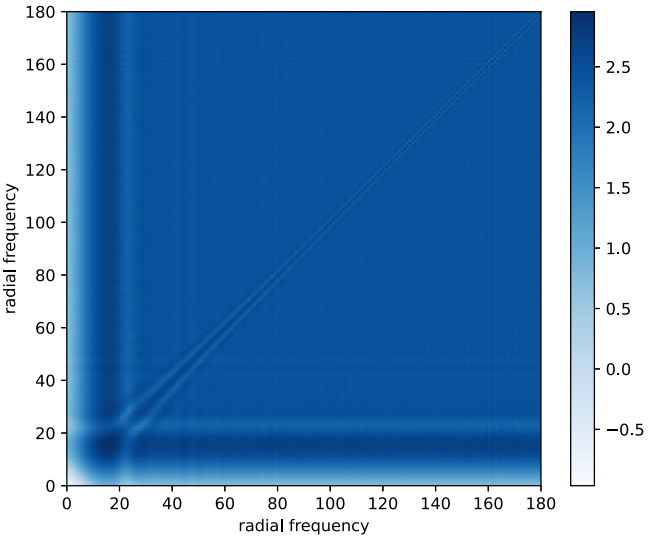

Figure 1. *We visualize* $\log_{10}\left(\sum_{i=1}^{N}\left|\widehat{h}_i(|\xi|)\right|^2|\widehat{h}_i(|\eta|)|^2\right)$ *for each pair of radial frequencies* $|\xi|,|\eta|$ *for the experimental dataset EMPIAR-10028*[18] *obtained from the Electron Microscopy Public Image Archive*[19]*. All values are greater than* -1 *in the log scale.*

where $\widehat{g}_i,\widehat{h}_i,\widehat{f}_i$, and $\widehat{\varepsilon}_i$ denote the Fourier transforms of g_i,h_i,f_i, and ε_i, respectively. Since taking the Fourier transform changes convolution to point-wise multiplication, if each Fourier transform $|\widehat{h}_i| \geq \delta$ for some $\delta > 0$, then we could estimate c by first estimating $\widehat{f}_i$ from $\widehat{g}_i$ for $i = 1,\ldots,N$, and then using the sample covariance matrix. However, in practice, the CTF $\widehat{h}_i$ is approximately a radial function with many zero-crossings, which means that multiplication by $\widehat{h}_i$ destroys information in the corresponding frequencies, making the restoration from a single image ill-posed.

2.3. Fourier–Bessel

The main ingredient in our fast covariance estimation is a fast transform into a convenient and computationally advantageous basis, known as the Fourier–Bessel basis (which consists of the harmonics on the disk: eigenfunctions of the Laplacian on the disk that obey Dirichlet boundary conditions). This specific choice of basis has a number of beneficial properties:

(i) it is orthonormal,
(ii) it is ordered by frequency,
(iii) it is steerable, that is, images can be rotated by applying a diagonal transform to the basis coefficients,
(iv) it is easy to convolve with radial functions, that is, images can be convolved with radial functions by applying a diagonal transform to the basis coefficients.

These properties have made the Fourier–Bessel basis a natural choice in a number of imaging applications[7,15,20–22] and will be central to the development of our fast covariance estimation method. In polar coordinates (r,θ) in the unit disk $\{(r,\theta) : r \in [0,1), \theta \in [0,2\pi)\}$, the Fourier–Bessel basis functions are defined by

$$\psi_{nk}(r,\theta) = \gamma_{nk} J_n(\lambda_{nk} r) e^{in\theta}, \tag{6}$$

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

where γ_{nk} is a normalization constant, J_n is the n-th order Bessel function of the first kind (see Ref. 23, Section 10.2), and λ_{nk} is its kth smallest positive root; the indices (n,k) run over $\mathbb{Z} \times \mathbb{Z}_{>0}$.

Recent work[16] has devised a new fast algorithm to expand $L \times L$-images into $\sim L^2$ Fourier–Bessel basis functions. Informally speaking, given $\sim L^2$ basis coefficients, the algorithm can evaluate the function on an $L \times L$ grid in $O(L^2 \log L)$ operations; the adjoint can be computed in the same number of operations, which makes iterative methods fast. Compared to previous expansion methods[7,22], it enjoys both theoretically guaranteed accuracy and lower time complexity.

2.4. Key property of Fourier–Bessel basis

A key property of the Fourier–Bessel basis is that convolution with radial functions is diagonal transformations in any truncated basis expansion. More precisely, the following result holds (Ref. 16, Lemma 2.3): suppose that $f = \sum_{(n,k) \in I} \alpha_{nk} \psi_{nk}$ for some index set I, and $h = h(|x|)$ is a radial function. Then,

$$P_{\mathcal{J}}(f * h) = \sum_{(n,k) \in I} \alpha_{nk} \widehat{h}(\lambda_{nk}) \psi_{nk}, \tag{7}$$

where $P_{\mathcal{J}}$ denotes orthogonal projection onto the span of $\{\psi_{nk}\}_{(n,k) \in I}$, $\widehat{h}$ is the Fourier transform of h, and λ_{nk} is the kth positive root of J_n.

We emphasize that the weights of the diagonal transform in Equation (7) are not the coefficients of h in the disk harmonic expansion. Indeed, since h is radial, it follows from Equation (6) that the coefficients β_{nk} of h in the basis ψ_{nk} satisfy $\beta_{nk} = 0$ when $n \neq 0$. Computing the weights $\widehat{h}(\lambda_{nk})$ from the coefficients β_{nk} of h in the basis, would involve computing weighted sums of the Fourier transforms of the basis functions: $\widehat{h}(\lambda_{nk}) = \sum_{l \in \mathcal{J}} \beta_{0l} \widehat{\psi}_{0l}(\lambda_{nk})$, for some index set $\mathcal{J}$.

As an alternative to the disk harmonic basis expansion, one can consider simply taking the Fourier transform of (1), which also leads to a diagonal representation of the convolution operator. However, the discrete Fourier transform does not have the steerability property, which is essential for the covariance estimation. Another attempt could be to use the polar Fourier transform. However, this representation is not invariant to arbitrary in-plane rotations, but only to finitely many rotations as determined by the discretization spacing of the grid. These expansions are therefore unsuitable for the goal of this article and we instead use expansions into the Fourier–Bessel basis, although other steerable bases could be considered[24,25]. Table 1 summarizes the considerations that make the Fourier–Bessel basis a natural choice of basis.

2.5. Block diagonal structure

The steerable property of the Fourier–Bessel basis implies that the 2D-covariance will be block diagonal in this basis. A full representation of an $L^2 \times L^2$ matrix requires $O(L^4)$ elements, but this is reduced to $O(L^3)$ nonzero entries by the block diagonal structure. This block diagonal structure follows from the

Table 1. *Summary of desirable properties of a few different basis candidates.*

Basis	Orth.	Cont. steerable	Radial convolution diag.	Fast expansion[a]
Real	✓	✗	✗	✓
2-D discrete Fourier	✓	✗	✓	✓
Polar Fourier	✗	✗ (discrete)	✓	✓
PSWF[24,25]	✓	✓	✗	✗
Fourier–Bessel[16][b]	✓	✓	✓	✓

[a]The basis expansion from Cartesian grid representation can be completed within $O(L^2)$ operations up to log factors.
[b]The new expansion algorithm[16] improves the previous computational method[7] in terms of accuracy guarantees, computational complexity, and the fact that it derives weights such that radial convolution is a diagonal operation.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

form of the basis functions and that the distribution of in-plane rotations is assumed to be uniform. Indeed, suppose that $f = \sum_{(n,k)\in I} \alpha_{nk}\psi_{nk}$. For simplicity assume that f has mean zero; subtracting the mean will only change the radial components, since the other components corresponding to nonvanishing angular frequencies have zero mean by merely averaging over all possible in-plane rotations. By Assumption A2, the covariance function in polar coordinates satisfies

$$c((r,\theta),(r',\theta')) = c((r,\theta+\varphi),(r',\theta'+\varphi)), \tag{8}$$

for all φ in $[0,2\pi]$. The covariance function in Equation (3) can be expanded in a double Fourier–Bessel basis expansion as

$$c((r,\theta),(r',\theta')) = \sum_{(n,k)\in I}\sum_{(n',k')\in I} C_{(nk,n'k')}\psi_{nk}(r,\theta)\overline{\psi_{n'k'}}(r',\theta'), \tag{9}$$

where $C_{(nk,n'k')}$ is the covariance matrix in the Fourier–Bessel basis. Combining Equations (8) and (9) and integrating φ over $[0,2\pi]$ gives

$$\begin{aligned} c(r,\theta),(r',\theta')) &= \frac{1}{2\pi}\int_0^{2\pi} c((r,\theta+\varphi),(r',\theta'+\varphi))d\varphi \\ &= \sum_{(n,k)\in I}\sum_{(n',k')\in I} C_{(nk,n'k')}\frac{1}{2\pi}\int_0^{2\pi}\psi_{nk}(r,\theta+\varphi)\overline{\psi_{n'k'}}(r',\theta'+\varphi)d\varphi \\ &= \sum_{(n,k)\in I}\sum_{(n',k')\in I} C_{(nk,n'k')}\psi_{nk}(r,\theta)\overline{\psi_{n'k'}}(r',\theta')\frac{1}{2\pi}\int_0^{2\pi} e^{i(n-n')\varphi}d\varphi \\ &= \sum_{(n,k)\in I}\sum_{(n',k')\in I} \delta_{n=n'}C_{(nk,n'k')}\psi_{nk}(r,\theta)\overline{\psi_{n'k'}}(r',\theta'), \end{aligned} \tag{10}$$

where $\delta_{n=n'}$ is a Dirac function that is equal to 1 if $n = n'$ and zero otherwise, and we note that the second to last equality uses the fact that $\psi_{nk}(r,\theta) = \gamma_{nk}J_n(\lambda_{nk}r)e^{in\theta}$. Since the coefficients $C_{(nk,n'k')}$ in the expansion Equation (9) are unique, it follows from Equation (10) that $C_{(nk,n'k')} = 0$ when $n \neq n'$. Hence, the covariance matrix $C_{(nk,n'k')}$ has a block diagonal structure whose blocks consist of the indices (nk,nk') for a given value of n. In the following section, we show how these properties enable a fast method to estimate the covariance matrix.

2.6. Covariance estimation

In the Fourier–Bessel basis, Equation (1) is written as

$$G_i = H_i \odot F_i + E_i, \tag{11}$$

where G_i, F_i, and E_i are coefficient vectors of g_i, f_i, ε_i in the Fourier–Bessel basis, respectively and H_i is the vector encoding the convolution operator of Section 2.4, that is, with components $\hat{h}_i(\lambda_{nk})$. The vectors are b-dimensional column vectors, where $b = O(L^2)$ is the number of basis coefficients. We use this simple structure to obtain a closed-form expression for the sample covariance matrix of the F_i. We estimate this matrix by minimizing the discrepancy between the sample covariance and the population covariance; more precisely, the estimated covariance matrix $\tilde{C}$ is computed by solving the least squares-problem

$$\tilde{C} = \arg\min_{C} \sum_{i=1}^{N} \left\| (G_i - H_i \odot \tilde{\mu})(G_i - H_i \odot \tilde{\mu})^T - \left(C \odot \left(H_i H_i^T\right) + \sigma^2 I\right)\right\|_F^2. \tag{12}$$

where $\tilde{\mu}$ is defined by

$$\tilde{\mu} = \arg\min_{\mu} \sum_{i=1}^{N} \|G_i - H_i \odot \mu\|^2, \tag{13}$$

whose solution is

$$\tilde{\mu} = \left(\sum_{i=1}^{N} H_i \odot G_i\right) \oslash \left(\sum_{i=1}^{N} H_i^{\odot 2}\right). \tag{14}$$

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

The least squares-solution of Equation (12) can be determined by the following system of linear equations

$$\left(\sum_{i=1}^{N} H_i^{\odot 2}\left(H_i^{\odot 2}\right)^T\right) \odot \tilde{C} = \sum_{i=1}^{N}\left(H_i H_i^T\right) \odot B_i - \sigma^2 \sum_{i=1}^{N} \mathrm{diag}\left(H_i^{\odot 2}\right), \tag{15}$$

where

$$B_i = (G_i - H_i \odot \tilde{\mu})(G_i - H_i \odot \tilde{\mu})^T. \tag{16}$$

It follows that

$$\tilde{C} = \left(\sum_{i=1}^{N}\left[B_i \odot \left(H_i H_i^T\right) - \sigma^2 \mathrm{diag}\left(H_i^{\odot 2}\right)\right]\right) \oslash \left(\sum_{i=1}^{N} H_i^{\odot 2}\left(H_i^{\odot 2}\right)^T\right). \tag{17}$$

As discussed in Section 2.5, the covariance matrix is block diagonal in the Fourier–Bessel basis. More precisely, the only nonzero elements $\tilde{C}(nk, n'k')$ of the matrix $\tilde{C}$ are those with $n = n'$. Therefore, the matrices in Equations (16) and (17) need only be calculated for this subset of indices. Since there is a total of $O(L^3)$ of these indices, this reduces the computational complexity compared to computing with the full matrices.

Note that the covariance matrix estimated from (Equation 17) may not be positive semidefinite due to subtraction of the term $\sigma^2 \mathrm{diag}\left(H_i^{\odot 2}\right)$. Therefore, when running the method in practice it is beneficial to use an eigenvalue shrinkage method. The computational cost of eigenvalue shrinkage for a matrix with our block structure is $O(L^4)$[15,26]. For completeness, we include this computational cost of eigenvalue shrinkage in our overall computational complexity. Informally speaking, the idea of eigenvalue shrinkage is to replace the term $\sum_{i=1}^{N}\left[B_i \odot \left(H_i H_i^T\right) - \sigma^2 \mathrm{diag}\left(H_i^{\odot 2}\right)\right]$ in Equation (17) by $\sum_{i=1}^{N}\left[B_i \odot \left(H_i H_i^T\right)\right]$, and then shrink and truncate the eigenvalues in a systematic way before Hadamard division[26]. The steps of the algorithm are summarized in Algorithm 1.

Algorithm 1 Fast covariance estimation method.

Input: Observed images g_i, radial functions h_i, $i = 1, \ldots, N$.
Output: Estimated covariance matrix $\tilde{C}$ in the domain of the Fourier–Bessel basis

1. Expand observed images g_i into the Fourier–Bessel basis, with resulting coefficient vector G_i
2. Compute the vectors H_i with components $\widehat{h_i}(\lambda_{nk})$, representing the action of the CTFs in the Fourier–Bessel basis
3. Compute sample mean $\tilde{\mu}$ from Equation (14)
4. Use Equation (16) to compute the elements $B_i(nk, nk')$ for all n, k, k'
5. Use Equation (17) to compute the elements of the sample covariance matrix $\tilde{C}(nk, nk')$ for all n, k, k', refined using eigenvalue shrinkage

The complexity of step 1 of the algorithm is $O(NL^2 \log L)$. The complexity of step 2 is $O(ML^2 \log L)$. The complexity of step 3 is $O(NL^2)$, since the number of basis coefficients is $O(L^2)$. The complexity of step 4 is $O(NL^3)$, since there are at most $O(L^3)$ nonzero elements with indices (nk, nk'). The complexity of step 5 is $O(NL^3 + L^4)$, where the additional term $\mathcal{O}(L^4)$ comes from the computational complexity of eigenvalue shrinkage. Thus, the total complexity of the algorithm is $O(NL^3 + L^4)$.

We now show how this can be used in an application to image denoising. Given an estimate of the covariance matrix $\tilde{C}$, the CWF approach estimates the F_i by a linear Wiener filter[15],

$$\tilde{F}_i = \tilde{\mu} + \tilde{C}\,\mathrm{diag}(H_i)\left(\tilde{C} \odot \left(H_i H_i^T\right) + \sigma^2 I\right)^{-1}(G_i - H_i \odot \tilde{\mu}), \tag{18}$$

See Ref. 15 for more details.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

3. Synthetic Data Results

We compare the timings of our fast method to previous approaches[15], for synthetic images generated from the 3-D volume of SARS-CoV-2 (Omicron) spike complexes[27] (EMD-32743), from the online EM data bank[28]. The original volume has size 512 in each dimension, with pixel size 0.832 Å. We downsample the original volume to size $L \times L \times L$, with $L = 32,128$ and 512, respectively, and show the computational times. To generate the synthetic noisy images, we first generate 10,000 clean projection images of the 3-D volume from random and uniformly distributed viewing directions. We next divide the set of clean images into a number of defocus groups, where the defocus values range from 1 μm to 4 μm. For all CTFs, we set the voltage as 300 kV and the spherical aberration as 2 mm. After convolving the images with their CTFs, we add colored noise with power spectral density $1/(rL/20+1)$ up to a constant scale, where $r \in [0,1]$ is the radial frequency. For both the previous method and ours, the CTFs and the noisy images are whitened before estimating the covariance. A few sample clean and noisy images are shown in Figure 5. All experiments were carried out on a machine with 750 GB memory and 72 Intel Xeon E7–8880 v3 CPUs running at 2.30GHz. We note that our implementation heavily relies on packages such as FINUFFT[29,30], NumPy[31], and SciPy[32], which are not fully optimized for parallel computing. The implementation of our fast method and the code of CWF[15] both effectively use only around 20 cores at runtime. Due to the complexity of modern computational architectures, performing a fair comparison can be challenging, but we have made every effort to do so. Both our code and the CWF code[15] take input images in a tensor format, and used vectorized operations whenever possible to avoid for-loops. The primary factor determining performance is not the specific implementation of the code, but rather the underlying computational complexities. Improving the parallelization is a technical direction for future work.

Figure 2 shows the time required to estimate the covariance matrices as a function of the number of defocus groups. Note that the runtime for the previous approach for the largest image and defocus group sizes is infeasibly large and that our fast method exhibits a speedup of up to three orders of magnitude.

Figure 3 shows the top six principal components estimated by our method and by traditional PCA using 10^4 raw images, where $L = 128$, SNR $= 0.1$ and $M = 100$ defocus groups, compared to traditional PCA on 10^6 images. We use the sample covariance matrix of phase-flipped images in real space for the traditional PCA. For all methods, we use λ to denote 100 times the eigenvalues of the eigenimages. The eigenimages from the traditional PCA look much noisier than ours, and contain artifacts that are due to imperfect CTF correction (see, e.g., the circular artifacts in top three eigenimages in Figure 3). The eigenimages from the traditional PCA also fail to preserve the symmetries (see, e.g., 6th eigenimage in Figure 3) that are present in our eigenimages, since they do not utilize the steerable basis and rotation-augmented images.

Figure 4 shows the quality of covariance estimation and image denoising when $L = 128$ and using $M = 100$ defocus groups. The quality of the covariance estimation is measured by the relative error in each angular frequency, which is defined as

$$\mathrm{err}_n = \|C_n - \tilde{C}_n\|_F / \|C_n\|_F, \tag{19}$$

where C_n and $\tilde{C}_n$ are respectively the nth diagonal blocks of the clean and estimated covariance matrix, corresponding to all indices of the form (nk, nk'). The clean covariance matrix was approximated by the sample covariance matrix of 10^6 clean projection images. The performance of image denoising is measured by the Fourier ring correlation (FRC) between the clean and denoised images. Namely, for the ith pair of clean and denoised images I_i^c and I_i^d, we first compute their Fourier coefficient vectors $\boldsymbol{f}_{i,r}^c, \boldsymbol{f}_{i,r}^d$ at radial frequency r by the nonuniform FFT[33–35], where $1 \le i \le N$ and $0 \le r \le r_{\max}$. We then compute their averaged correlation for each r:

$$\mathrm{FRC}(r) = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle \boldsymbol{f}_{i,r}^c, \boldsymbol{f}_{i,r}^d \rangle}{\|\boldsymbol{f}_{i,r}^c\| \|\boldsymbol{f}_{i,r}^d\|},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two complex vectors. The FRC is a real-valued quantity due to a symmetry property that arises since the images I_i^c, I_i^d are real-valued.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

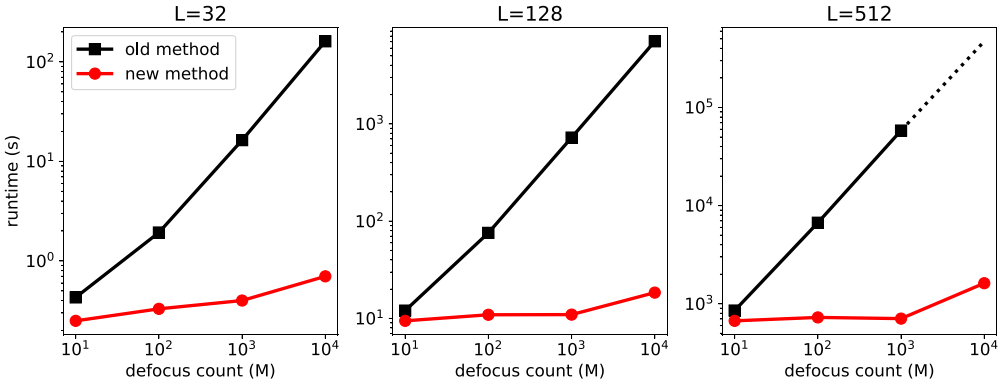

Figure 2. *Timing comparison for covariance matrix estimation of 10,000 images of size $L \times L$. The old method[15] timing for $L = 512$ and defocus count 10^4 is extrapolated due to time and memory constraints.*

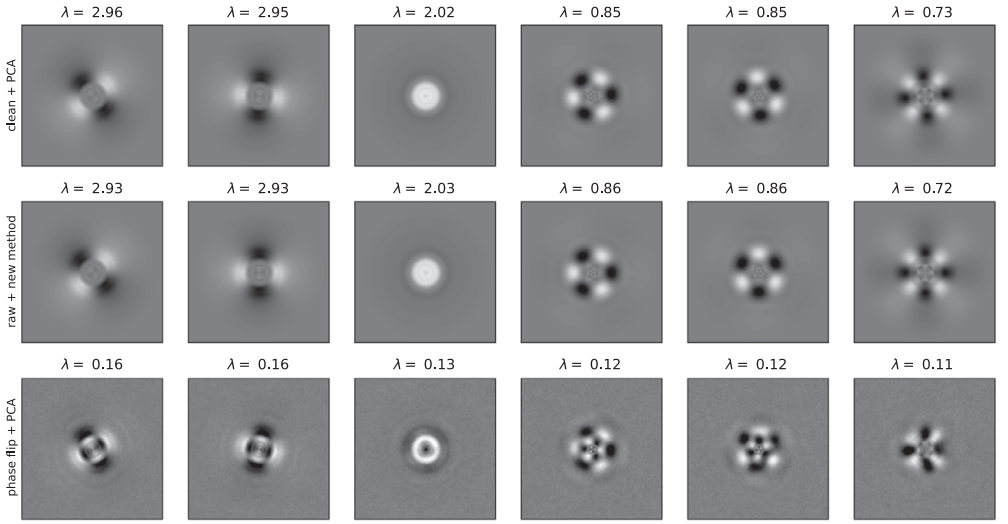

Figure 3. *Top six eigenimages computed by traditional PCA on 10^6 clean images (top panel), our new method on 10^4 raw images (middle panel), and traditional PCA on 10^4 phase-flipped images (bottom panel). The signal-to-noise ratio for the images for the new method and traditional PCA was 0.1.*

In addition to the speedups of Figure 2, the left panel of Figure 4 demonstrates a slight increase in the estimation quality of our proposed algorithm, compared to the previous approach. This is possibly caused by improved accuracy in the Fourier–Bessel basis approximation as well as improved accuracy by using the closed-form expression Equation (17) compared to the approximate conjugate gradient step of previous approaches. Similarly, on the right panel, the Fourier ring correlation between the denoised and clean images shows a slight performance increase. Figure 5 shows sample denoised images for different values of the SNR where $L = 128$ and $M = 100$. As a comparison, we show images denoised using the approach of this paper and the CWF method[15].

4. Experimental Data Results

We conclude by using our method on two experimental datasets obtained from the Electron Microscopy Public Image Archive[19], namely EMPIAR-10028[18] and EMPIAR-10081[36]. EMPIAR-10028 is a

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

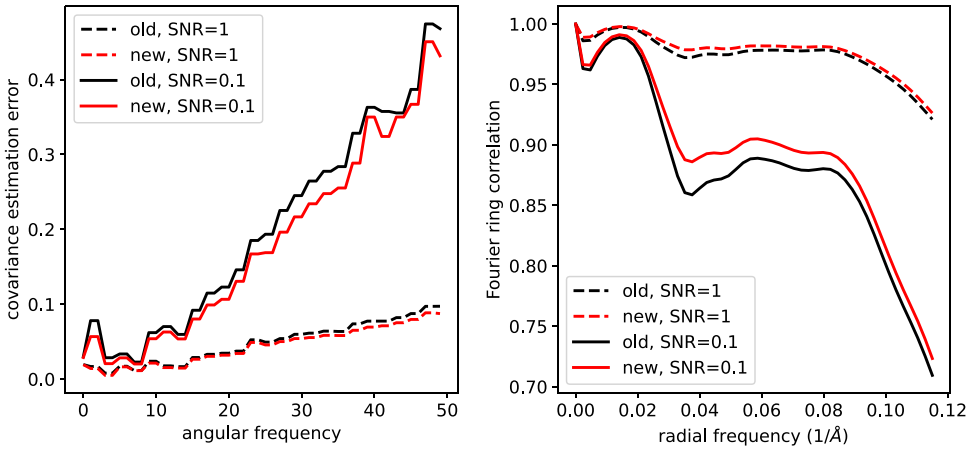

Figure 4. *Relative estimation error of the covariance matrix (left) and the Fourier ring correlation between the denoised and clean images (right).*

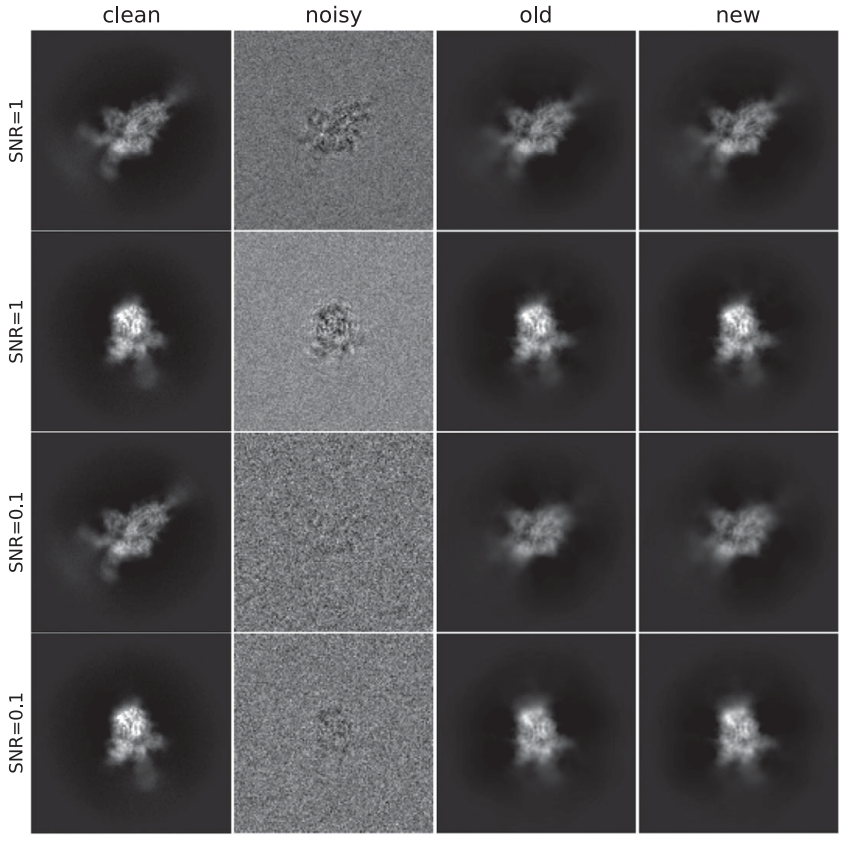

Figure 5. *Clean, noisy and denoised images. The covariance estimation used* $N = 10,000$ *images, and parameters* $L = 128$ *and* $M = 100$.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

Table 2. *Timing comparison in seconds for EMPIAR-10028 (top) and EMPIAR-10081 (bottom).*

EMPIAR-10028					
Methods	T_{ffb}	T_{ctf}	T_{cov}	T_{denoise}	T_{total}
Old CWF, $L = 360$	1415	598	27550	201	29764
Fast CWF, $L = 360$	768	5	2220	95	3088
EMPIAR-10081					
Methods	T_{ffb}	T_{ctf}	T_{cov}	T_{denoise}	T_{total}
Old CWF, $L = 128$	47	3369	46318	45	49779
Fast CWF, $L = 128$	35	8	93	27	163
Old CWF, $L = 256$	363	NA	NA	NA	NA
Fast CWF, $L = 256$	169	34	1007	163	1373

Note. For EMPIAR-10081, when $L = 256$, the old CWF method[15] encounters memory issues and cannot be run until completion. T_{ffb}, the time required to expand all images in the Fourier–Bessel basis (step 1 in Algorithm 1); T_{ctf}, the time to compute a matrix representation of the application of the point spread function (step 2 in Algorithm 1); T_{cov}, the time to estimate the covariance matrix (steps 3–5 in Algorithm 1); T_{denoise}, the time to denoise the number of images indicated in the main text (2014 images for EMPIAR-10028 and 502 images for EMPIAR-10081); T_{total}, the total computational time.

dataset of the Plasmodium falciparum 80S ribosome bound to the antiprotozoan drug emetine whose 3-D reconstruction is available in the EM data bank as EMD-2660[18]. The dataset contains 105247 motion corrected and picked particle images, from 1081 defocus groups, of size 360×360 with 1.34 Å pixel size. EMPIAR-10081 is a dataset of the human HCN1 hyperpolarization-activated cyclic nucleotide-gated ion channel, whose 3-D reconstruction can be found in the EM data bank as EMD-8511[36]. The dataset contains 55,870 motion corrected and picked particle images, from 53,384 defocus groups, of size 256×256 with 1.3 Å pixel size.

Computational times are shown in Table 2, showing a speedup of more than two orders of magnitude for the datasets with the largest number of distinct CTFs. Note that regular CWF on EMPIAR-10081 encounters memory issues and cannot be run to completion, whereas our fast method runs seamlessly. The old CWF has much higher space complexity $O(ML^3)$ for covariance estimation than that of our method $O(L^3)$. The reason is that the old CWF represents CTFs using block diagonal matrices (each takes $O(L^3)$ memory). Moreover, the old CWF loads all these block diagonal CTFs at once in memory so as to define the linear operator for the conjugate gradient method. Our method circumvents the need to store all CTFs at once due to our closed-form solution, and therefore we can read CTFs in batches. Moreover, each CTF has diagonal representation in our method, which means much lower space complexity $O(L^2)$. The $O(L^3)$ space complexity for our method is only for storing a single covariance matrix.

For EMPIAR-10081, storing the CTFs for the old method takes approximately $50{,}000 \times (256^3/6) \times 8/10^9 \approx 1118$ GB which is above the 750 GB memory limit of our machine. For the new method, storing the covariance matrix takes $(256^3/6) \times 8/10^9 \approx 0.02$ GB and storing the CTFs takes $1000 \times 256^2 \times 8/10^9 \approx 0.52$ GB (with batch size 1000) that can be even handled by a common laptop.

In order to obtain a comparison, we therefore additionally downsample these images to $L = 128$ where the original CWF can successfully run. On EMPIAR-10028, we used all images for covariance estimation, and denoised the 2014 images from 0 th, 50 th, 100 th, …, 1000 th defocus groups. On EMPIAR-10081, we used all images for covariance estimation, and denoised the 502 images from 0th, 100th, 200th, …, 50,000th defocus groups. For both old and new methods, the covariance matrices are further refined to correct image contrast variations[21]. Sample visualization results are shown in Figures 6 and 7. The old CWF method cannot handle the full resolution recovery for EMPIAR-10081 (with about

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

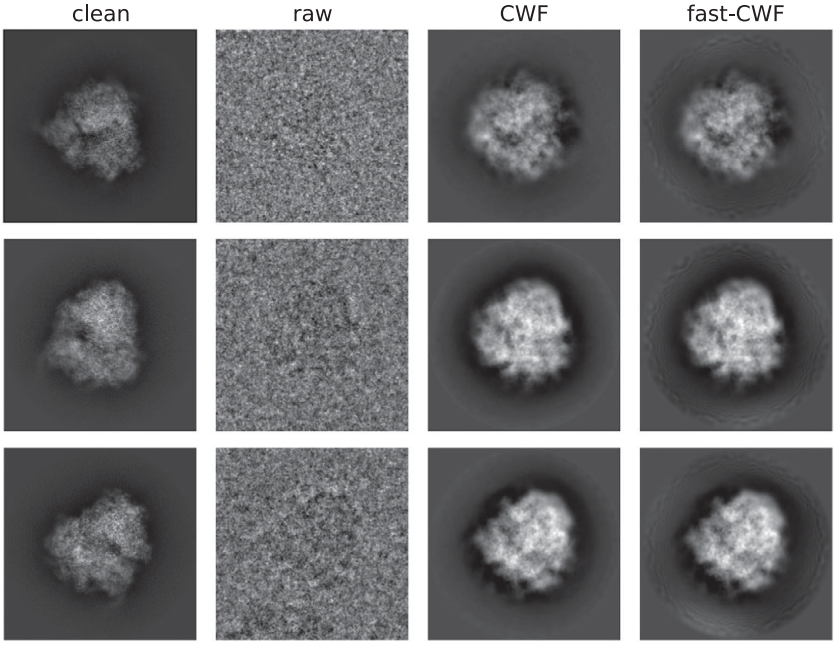

Figure 6. *Denoised images (EMPIAR-10028). The method used* $N = 105{,}247$ *images, from* $M = 1081$ *defocus groups, of size* 360×360*. The clean images are obtained by aligning* 1000 *clean projection images (from uniformly distributed viewing directions) with phase-flipped raw images.*

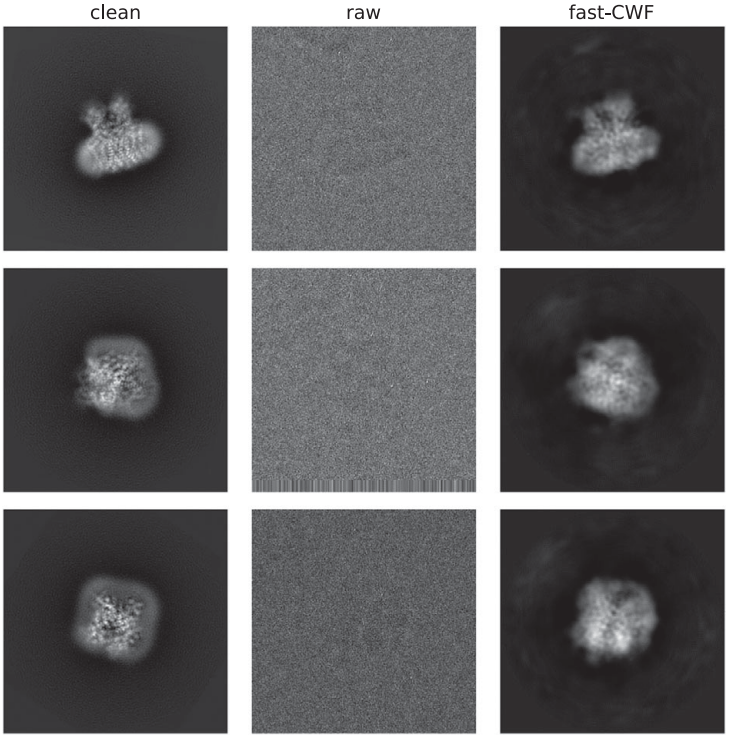

Figure 7. *Denoised images (EMPIAR-10081). The method used* $N = 55{,}870$ *images, from* $M = 53{,}384$ *defocus groups, of size* 256×256*. The clean images are obtained by aligning 1000 clean projection images (from uniformly distributed viewing directions) with phase-flipped raw images.*

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

50,000 distinct CTFs) due to its high space complexity. Moreover, the old CWF method does not return satisfying restored images for downsampled low resolution images. The reason is that for EMPIAR-10081, almost all images belong to different defocus groups, and the estimated noise power spectrum from a single image is too noisy. Therefore, EMPIAR-10081 is a failure case for the old CWF method in terms of both accuracy and computation time. In contrast, our method assumes radially symmetric noise power spectrum and estimates it using NUFFT over concentric rings with different radii, and averages over each ring. This averaging process largely reduces the noise in the estimated power spectrum.

5. Discussion

Covariance estimation and PCA of cryo-EM images are key ingredients in many classic cryo-EM methods including multivariate statistical analysis[3–6] and Kam's method for ab initio modeling[8]. We propose a fast method to estimate the covariance matrix of noisy cryo-EM images, and then illustrate its application to simultaneously correct for the CTFs and denoise the images. The approach relies on recent improvements to algorithms for expanding images in the Fourier–Bessel basis[16], and has time complexity $O\left(NL^3 + L^4\right)$ which is independent of the number of defocus groups. Our new approach is both significantly faster and more memory-efficient compared to the previous CWF method[15] and we apply our method to large experimental datasets with many distinct CTFs with speedups by factors up to more than two orders of magnitude. Our approach could potentially be extended to higher-dimensional data and to the setting where images are distorted by CTFs which are not exactly radial, using either analytical correction terms or iterative numerical steps.

Acknowledgments. We are grateful to Kenny Huang for running some numerical experiments in the initial stage of this project.

Competing Interests. The authors declare no competing interests exist.

Authorship Contributions. N.F.M., O.M., Y.S., and A.S. conceived the project. N.F.M., O.M., and Y.S. designed the algorithms. Y.S. wrote the software and performed the experiments. All authors wrote the article and approved the final submission.

Funding Statement. N.F.M. was supported in part by NSF DMS-1903015 and was visiting the Institute for Pure and Applied Mathematics (IPAM) when part of this research was performed, which was supported by NSF DMS-1925919. A.S. was supported by grants from the AFOSR FA9550-20-1-0266; the Simons Foundation Math+X Investigator Award; NSF BIGDATA Award IIS-1837992; NSF DMS-2009753; and NIH/NIGMS 1R01GM136780–01.

Data Availability Statement. Replication data and code can be found at https://github.com/yunpeng-shi/fast-cryoEM-PCA.

References

1. Liu W & Frank J (1995) Estimation of variance distribution in three-dimensional reconstruction. i. Theory. *JOSA A* **12**(12), 2615–2627.
2. Penczek PA, Kimmel M & Spahn CM (2011) Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-em images. *Structure* **19**(11), 1582–1590.
3. Van Heel M (1984) Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy* **13**(1–2), 165–183.
4. Van Heel M & Frank J (1981) Use of multivariates statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**(1), 187–194.
5. van Heel M, Portugal R **and Schatz M** (2009) *Multivariate Statistical Analysis in Single Particle (Cryo) Electron Microscopy. An Electronic Textbook: Electron Microscopy in Life Science. 3D-EM Network of Excellence*. London.
6. van Heel M, Portugal RV, Schatz M, *et al.* (2016) Multivariate statistical analysis of large datasets: Single particle electron microscopy. *Open J Stat* **6**(04), 701.
7. Zhao Z & Singer A (2014) Rotationally invariant image representation for viewing direction classification in cryo-EM. *J Struct Biol* **186**(1), 153–166.
8. Kam Z (1980) The reconstruction of structure from electron micrographs of randomly oriented particles. *J Theor Biol* **82**(1), 15–39.
9. Sharon N, Kileel J, Khoo Y, Landa B & Singer A (2020) Method of moments for 3D single particle ab initio modeling with non-uniform distribution of viewing angles. *Inverse Probl* **36**(4), 044003.

https://doi.org/10.1017/S2633903X23000028 Published online by Cambridge University Press

10. Bendory T, Khoo Y, Kileel J, Mickelin O & Singer A (2022) Autocorrelation analysis for cryo-EM with sparsity constraints: improved sample complexity and projection-based algorithms. Preprint, arXiv:2209.10531.
11. Huang S, Zehni M, Dokmanić I & Zhao Z (2022) Orthogonal matrix retrieval with spatial consensus for 3D unknown-view tomography. Preprint, arXiv:2207.02985.
12. Levin E, Bendory T, Boumal N, Kileel J & Singer A (2018) 3D ab initio modeling in cryo-EM by autocorrelation analysis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 1569–1573. IEEE.
13. Bhamre T, Zhang T & Singer A (2015) Orthogonal matrix retrieval in cryo-electron microscopy. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 1048–1052. IEEE.
14. Bhamre T, Zhang T & Singer A (2022) Anisotropic twicing for single particle reconstruction using autocorrelation analysis. U. S. Patent No. 11,227,403. arXiv:1704.07969.
15. Bhamre T, Zhang T & Singer A (2016) Denoising and covariance estimation of single particle cryo-EM images. *J Struct Biol* **195**(1), 72–81.
16. Marshall NF, Mickelin O & Singer A (2022) Fast expansion into harmonics on the disk: a steerable basis with fast radial convolutions. Preprint, arXiv:2207.13674.
17. Bandeira AS, Blum-Smith B, Kileel J, Perry A, Weed J & Wein AS (2017) Estimation under group actions: recovering orbits from invariants. Preprint, arXiv:1712.10163.
18. Wong W, Bai X-c, Brown A, *et al.* (2014) Cryo-EM structure of the plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3**, e03080.
19. Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ & Patwardhan A (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* **13**(5), 387–388.
20. Rangan A, Spivak M, Andén J & Barnett A (2020) Factorization of the translation kernel for fast rigid image alignment. *Inverse Probl* **36**(2), 024001.
21. Shi Y & Singer A (2022) Ab-initio contrast estimation and denoising of Cryo-EM images. *Comput Methods Programs Biomed* **224**, 107018.
22. Zhao Z, Shkolnisky Y & Singer A (2016) Fast steerable principal component analysis. *IEEE Trans Comput Imaging* **2**(1), 1–12.
23. Lozier D (2003) *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.1.5 of 2022-03-15.
24. Landa B & Shkolnisky Y (2017) Approximation scheme for essentially bandlimited and space-concentrated functions on a disk. *Appl Comput Harmon Anal* **43**(3), 381–403.
25. Landa B & Shkolnisky Y (2017) Steerable principal components for space-frequency localized images. *SIAM J Imaging Sci* **10** (2), 508–534.
26. Donoho DL, Gavish M & Johnstone IM (2018) Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann Stat* **46** (4), 1742.
27. Guo H, Gao Y, Li T, *et al.* (2022) Structures of omicron spike complexes and implications for neutralizing antibody development. *Cell Reports* **39**(5), 110770.
28. Lawson CL, Patwardhan A, Baker ML, *et al.* (2016) EMDataBank unified data resource for 3DEM. *Nuc Acids Res* **44**(D1), D396–D403.
29. Barnett AH (2021) Aliasing error of the $\exp\left(\beta\sqrt{1-z^2}\right)$ kernel in the nonuniform fast Fourier transform. *Appl Comput Harmon Anal* **51**, 1–16.
30. Barnett AH, Magland J & Klinteberg L (2019) A parallel nonuniform fast Fourier transform library based on an "exponential of semicircle" kernel. *SIAM J Sci Comput* **41**(5), C479–C504.
31. Harris CR, Millman KJ, van der Walt SJ, *et al.* (2020) Array programming with NumPy. *Nature* **585**(7825), 357–362.
32. Virtanen P, Gommers R, Oliphant TE, *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* **17**, 261–272.
33. Dutt A & Rokhlin V (1993) Fast Fourier transforms for nonequispaced data. *SIAM J Sci Comput* **14**(6), 1368–1393.
34. Greengard L & Lee J-Y (2004) Accelerating the nonuniform fast Fourier transform. *SIAM Review* **46**(3), 443–454.
35. Lee J-Y & Greengard L (2005) The type 3 nonuniform FFT and its applications. *J Comput Phys* **206**(1), 1–5.
36. Lee C-H & MacKinnon R (2017) Structures of the human HCN1 hyperpolarization-activated channel. *Cell* **168**(1–2), 111–120.

Cite this article: Marshall N. F, Mickelin O, Shi Y and Singer A (2023). Fast principal component analysis for cryo-electron microscopy images. *Biological Imaging*, 3: e2. doi:https://doi.org/10.1017/S2633903X23000028