MAHALANOBIS DISTANCE FOR CLASS AVERAGING OF CRYO-EM IMAGES

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

Single particle reconstruction (SPR) from cryo-electron microscopy (EM) is a technique in which the 3D structure of a molecule needs to be determined from its CTF-affected, noisy 2D projection images taken at unknown viewing directions. 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 and is being used for selecting particles from micrographs and for inspecting the particle images [1]. We introduce a new metric, akin to the Mahalanobis distance, to compare cryo-EM images belonging to different defocus groups. The new metric is employed to rank nearest neighbors, thereby leading to an improved algorithm for class averaging. We evaluate the performance of the improved class averaging on synthetic datasets and note an improvement compared to [1].

***Index Terms*—** Cryo-electron microscopy, single particle reconstruction, particle picking, class averaging, Mahalanobis distance, denoising, CTF.

1. INTRODUCTION

SPR from cryo-EM is a rapidly advancing technique in structural biology to determine the 3D structures of macromolecular complexes in their native state, without the need for crystallization. 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 acquired raw projection images, the preprocessed images are next classified, and images within each class are averaged (known in the cryo-EM community as "class averaging"), to obtain a single image per class, that enjoys a higher signal to noise ratio (SNR) than the individual images. 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 [2]. 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. A second round of class averaging needs to be performed to identify and discard outliers after this step. The resulting class averages enjoy a much higher SNR than the input raw images, thereby allowing inspection of the dataset and elimination of outliers. Class averages are used for subsequent stages of the SPR pipelines, such as viewing angle detection, and finally, determination of the 3D structure.

ADD BACKGROUND WORK CLASS AVG

In [1], the authors introduced a new rotationally invariant representation to compute the distance between pairs of cryo-EM images. The images are first expanded in a steerable basis. Next, a rotationally invariant representation of the images is computed from this expansion. The rotationally invariant features are built by constructing the bispectrum from the expansion coefficients of the images in the steerable basis. After projecting the computed bispectrum to a lower dimensional space using a randomized PCA algorithm, the distance between images is computed as the Euclidean distance between their rotationally invariant representation. This completes the initial round of 2D classification.

The nearest neighbors detected at this stage can suffer from many outliers at high levels of noise. The initial classification is next improved by checking the consistency of in-plane rotations along several paths between nearest neighbors, by employing Vector Diffusion Maps (VDM)[3]. Finally, the detected nearest neighbors are aligned in-plane and averaged.

Recently [4], it was shown that this preliminary inspection can in fact 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 observations. This estimated covariance is then used in the classical Wiener filtering framework to obtain denoised images, which can be used for a preliminary viewing of the underlying dataset, and outlier detection.

There are two main contributions of this paper. First, we introduce a new metric, which can be viewed as a Mahalanobis distance [5], to compute the distance between pairs of

Thanks to XYZ agency for funding.

cryo-EM images. Second, we use the proposed Mahalanobis distance to improve the class averaging algorithm described in [1]. We first obtain for each image a list of S other images suspected as nearest neighbors using the algorithm described above (see section 2 for details), and then rank these suspects using the Mahalanobis distance. 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 synthetic datasets and observe an increase in the number of nearest neighbors correctly detected.

2. BACKGROUND

2.1. Image Formation Model

Under the linear, weak phase approximation [6], the image formation model in cryo-EM is given by

$$y_i = a_i \star x_i + n_i \tag{1}$$

where and $\star$ 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 contrast transfer function (CTF), and eqn.(1) can be rewritten as

$$Y_i = A_i X_i + N_i \tag{2}$$

The CTF is mathematically given by [7]

$$CTF(\hat{k}; \Delta\hat{z}^2) = \sin[-\pi\Delta\hat{z}\hat{k}^2 + \frac{\pi}{2}\hat{k}^4] \tag{3}$$

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". CTF's corresponding to different defocus values have different zero crossings (see Fig.1).

2.2. Class Averaging

The procedure for class averaging, described in [1], consists of 3 main steps. First, Fourier Bessel Steerable PCA is used to denoise images and estimate the 2D covariance matrix of the images and their in-plane rotations. Phase flipping is used as a CTF correction scheme. Second, the bispectrum of the expansion coefficients in the steerable basis is computed. The bispectrum is a rotationally invariant representation of images, but is typically of very high dimensionality. It is projected on to a lower dimensional subspace using a fast, randomized PCA algorithm. One way to compare the distance between images after this step is to use the normalized cross correlation. However, this method suffers from outliers in the nearest neighbor detection at very low SNR's. So, the third step uses Vector Diffusion Maps to improve the initial classification by viewing angles.

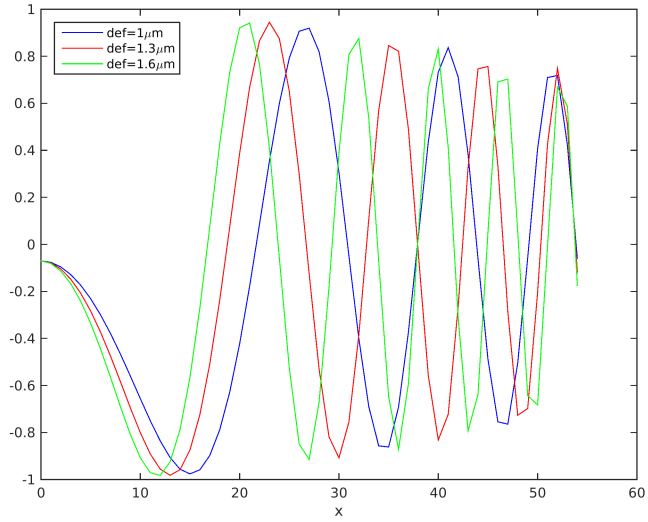

Fig. 1: CTF's for different values of the defocus.

2.3. Covariance Wiener Filtering (CWF)

CWF was proposed in [4] as an algorithm to (i) estimate the CTF-corrected covariance matrix of the underlying clean 2D projection images and (ii) using the estimated covariance to solve the associated deconvolution problem in 2 to obtain denoised images, that are estimates of X_i for each i in 2. The first step involves estimating the mean image of the dataset, μ, as $\hat{\mu}$, followed by solving a least squares problem to estimate the covariance Σ as $\hat{\Sigma}$. Under the assumption of additive white Gaussian noise, the estimate of the underlying clean image X_i is given by

$$\hat{X_i} = (I - H_i A_i)\hat{\mu} + H_i Y_i \tag{4}$$

3. MAHALONOBIS DISTANCE

Our goal is to define a metric to compare how close a any two cryo-EM images are, that is, given the CTF-affected, noisy observations for a pair of images

$$Y_i = H_i X_i + N_i, \quad \text{for} \quad i = 1, 2. \tag{5}$$

Due to the numerous zero crossings of the CTF, CTF correction is a challenging problem. A popular albeit heuristic approach for CTF correction is 'phase flipping', which involves simply inverting the sign of the Fourier coefficients. This corrects for the phase inversion caused due to the CTF, but not for amplitude correction. In [4], the authors introduced a new approach for denoising and CTF correction in a single step, called CWF. When comparing distances between cryo-EM images, one must take into account that different images belong to different defocus groups, that is, they are affected

by different CTF's. 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 also suffer from incorrectly identified neighbors. The main motivation of this paper is to introduce a Mahalanobis distance for cryo-EM images, as a metric to compute the distance between images belonging to different defocus groups. Moreover, we propose to use this notion of distance to improve the existing class averaging pipeline in [8].

We define a metric that gives a notion of distance between two underlying clean images X_1 and X_2. In our statistical model, the underlying clean images $X_1, X_2, \dots X_n$ (where n is the total number of images) 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.

$$\begin{aligned} X_1, X_2 &\sim N(\mu, \Sigma) \\ N_1, N_2 &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{6}$$

We denote the covariance of Y_1 and Y_2 by K_1 and K_2.

$$Cov(Y_i) = H_i \Sigma {H_i}^T + \sigma^2 I_n = K_i, \quad \text{for} \quad i = 1, 2 \tag{7}$$

Using the Guassian property, we have the following probability density functions (pdf)

$$f_{X_i}(x_i) = P \exp\{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\}, \quad \text{for} \quad i = 1, 2 \tag{8}$$

$$f_N(z_i) = Q \exp\{-\frac{1}{2}(z_i)^T \frac{1}{\sigma^2}(z_i)\}, z_i = y_i - H_1 x_i, \quad \text{for} \quad i = 1, 2 \tag{9}$$

$$f_{Y_i}(y_i) = R \exp\{-\frac{1}{2}(y_i - H_i\mu)^T {K_1}^{-1}(y_i - H_i\mu)\}, \quad \text{for} \quad i = 1, 2 \tag{10}$$

where $P = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}$, $Q = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n}$, and $R = \frac{1}{(2\pi)^{\frac{n}{2}} |K_1|^{\frac{1}{2}}}$.

$$\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ H_1 & I \end{bmatrix} \times \begin{bmatrix} X_1 \\ N_1 \end{bmatrix} \tag{11}$$

$$\sim N\left[\begin{bmatrix} \mu \\ H_1\mu \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma H_1^T \\ H_1\Sigma & H_1 \Sigma H_1^T + \sigma^2 I \end{bmatrix}\right] \tag{12}$$

Using conditional distributions

$$f_{X_1|Y_1}(x_1|y_1) \sim N(\alpha, L) \tag{13}$$

$$f_{X_2|Y_2}(x_2|y_2) \sim N(\beta, M) \tag{14}$$

where

$$\begin{aligned} \alpha &= \mu + \Sigma H_1^T (H_1 \Sigma H_1^T + \sigma^2 I)^{-1}(y_1 - H_1\mu) \\ L &= \Sigma - \Sigma H_1^T (H_1 \Sigma H_1^T + \sigma^2 I)^{-1} H_1 \Sigma \\ \beta &= \mu + \Sigma H_2^T (H_2 \Sigma H_2^T + \sigma^2 I)^{-1}(y_2 - H_2\mu) \\ M &= \Sigma - \Sigma H_2^T (H_2 \Sigma H_2^T + \sigma^2 I)^{-1} H_2 \Sigma \end{aligned} . \tag{15}$$

So

$$P(x_1 - x_2 | y_1, y_2) \sim N(\alpha - \beta, L + M) \tag{16}$$

Let $x_1 - x_2 = x_3$, and $\alpha - \beta = \gamma$. Then

$$\begin{aligned} &P(||x_3||_\infty < \epsilon | y_1, y_2) = P(||x_3||_\infty < \epsilon | y_1, y_2) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |L + M|^{\frac{1}{2}}} \times \\ &\int_{-\epsilon}^{\epsilon} \exp\{-\frac{1}{2}(x_3 - \gamma)^T (L + M)^{-1}(x_3 - \gamma\} dx_3 \end{aligned} \tag{17}$$

For small ϵ this is

$$= \frac{(2\epsilon)^n}{(2\pi)^{\frac{n}{2}} |L + M|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\gamma^T (L + M)^{-1}\gamma\} \tag{18}$$

So we can define our metric after taking the logarithm on both sides of eqn.(18), dropping out the constant term, and substituting back γ

$$-\frac{1}{2}\log(|L + M|) - \frac{1}{2}(\alpha - \beta)^T (L + M)^{-1}(\alpha - \beta) \tag{19}$$

4. ALGORITHM FOR IMPROVED CLASS AVERAGING USING MAHALANOBIS DISTANCE

We propose an improved class averaging algorithm that incorporates the Mahalanobis distance. The quantities α, β, L, M are computed for each image and defocus group respectively, using CWF [4]. The estimated covariance using CWF is block diagonal in the Fourier Bessel basis. In practice, we use α, β, L, M projected onto the subspace spanned by the principal components (for angular frequency block). We obtain an initial list of S nearest neighbors for each image using the Initial Classification algorithm in [8]. Then, for the list of nearest neighbors corresponding to each image, the Mahalanobis distance is computed and used to pick the closest K nearest neighbors, where $K < S$. The details of the algorithm are listed in Algorithm 1.

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 10000 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 shoud 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 neighbors, the Mahalanobis distance was computed between each image and its 50 aligned nearest neighbors. We then pick the closest $K = 10$ neighbors for

Algorithm 1 Improved Class Averaging

procedure INITIAL CLASSIFICATION [8]
    Image compression and denoising: compute Fourier Bessel steerable basis for images [9]
    Rotationally invariant features: compute the bispectrum from denoised coefficients in the steerable basis
    Randomized PCA[10] of high dimensional feature vectors from the bispectrum
    Initial nearest neighbor classification and alignment using brute force or fast randomized approximate nearest neighbor search [11]
    (Optional) Improve nearest neighbor classification using Vector Diffusion Maps (VDM) [3]
    Obtain a list of S nearest neighbors for each image
procedure CLASSIFICATION USING MAHALONOBIS DISTANCE
    Compute the quantities α, β, L, M using Covariance Wiener Filtering (CWF) [4]
    For each image and its S aligned nearest neighbors, compute the Mahalanobis distance between the image and neighbors
    Rank S neighbors according to the Mahalanobis distance, and choose the top K as nearest neighbors

Table 1: Number of nearest neighbors with correlation > 0.9, using 10000 images, $K = 10$ and $S = 50$.

	VDM		No VDM	
SNR	New	Old	New	Old
1/60	34965	32113	34537	29219
1/100	17262	14431	16057	13706

each image. For comparison, we compute 10 nearest neighbors for each image using only Initial Classification (with or without using the optional VDM step). Table 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 quality of nearest neighbors detected by the improved classification algorithm using the Mahalanobis distance. Figure 2 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.

6. FUTURE WORK

The Mahalonobis distance proposed here can be used as a metric for any manifold learning procedure [12, 13] such as diffusion maps [3, 14], with or without missing data.

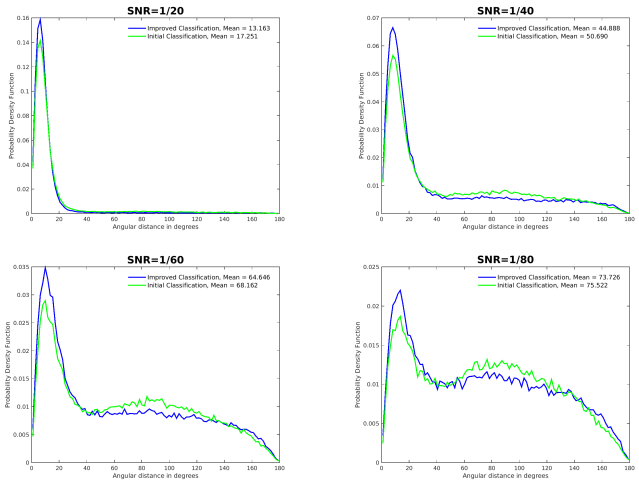

Fig. 2: The estimated probability density function of the angular distance (in degrees) between images classified into the same class by 1) Initial Classification and 2) Improved Classification using the Mahalanobis distance at different SNR's.

7. CONCLUSION

8. REFERENCES

[1] Z. Zhao and A. Singer, "Rotationally invariant image representation for viewing direction classification in cryo-EM," *Journal of Structural Biology*, vol. 186, no. 1, pp. 153 – 166, 2014.

[2] Sjors H.W. Scheres, "Semi-automated selection of cryo-em particles in relion-1.3," *Journal of Structural Biology*, vol. 189, no. 2, pp. 114 – 122, 2015.

[3] A. Singer and H.-T. Wu, "Vector diffusion maps and the connection laplacian," *Communications on Pure and Applied Mathematics*, vol. 65, no. 8, pp. 1067–1144, 2012.

[4] Tejal Bhamre, Teng Zhang, and Amit Singer, "Denoising and covariance estimation of single particle cryo-em images," *Journal of Structural Biology*, vol. 195, no. 1, pp. 72 – 81, 2016.

[5] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, Apr. 1936, vol. 2, pp. 49–55.

[6] Joachim Frank, "Chapter 2 - electron microscopy of macromolecular assemblies," in *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, Joachim Frank, Ed., pp. 12 – 53. Academic Press, Burlington, 1996.

[7] Joachim Frank, "Chapter 3 - electron microscopy of macromolecular assemblies," in *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, Joachim Frank, Ed., p. 43. Academic Press, Burlington, 1996.

[8] Z. Zhao and A. Singer, "Fourier Bessel rotational invariant eigenimages," *J. Opt. Soc. Am. A*, vol. 30, no. 5, pp. 871–877, May 2013.

[9] Z. Zhao, Y. Shkolnisky, and A. Singer, "Fast steerable principal component analysis," *IEEE Transactions on Computational Imaging*, vol. 2, no. 1, pp. 1–12, March 2016.

[10] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert, "A randomized algorithm for principal component analysis," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1100–1124, 2010.

[11] Peter Wilcox Jones, Andrei Osipov, and Vladimir Rokhlin, "Randomized approximate nearest neighbors algorithm," *Proceedings of the National Academy of Sciences*, vol. 108, no. 38, pp. 15679–15686, 2011.

[12] Ronen Talmon and Ronald R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12535–12540, 2013.

[13] Amit Singer and Ronald R. Coifman, "Non-linear independent component analysis with diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 25, no. 2, pp. 226 – 239, 2008.

[14] Ronald R. Coifman and Stphane Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5 – 30, 2006.