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 contrast transfer function (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 is used for selecting particles from micrographs and for inspecting the particle images. 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.

***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 [1, 2], 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 sleetced raw particle images, the images are next classified, and images within each class are averaged, (a step known as "class averaging"), to obtain a single image per class, that enjoys a higher 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 [3]. 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 orientation estimation, and finally, determination of the 3D structure.

The two popular approaches for 2D class averaging [4, 5, 6, 7] in cryo-EM are multivariate statistical analysis (MSA)[7] with multi-reference alignment (MRA) [8] and iterative reference-free alignment using K-means clustering [5]. TODO

In [9], the authors introduced a new, fast approach for 2D class averaging, based on a new rotationally invariant representation to compute the similarity between pairs of cryo-EM images.

Recently in [10], it was shown that this preliminary inspection of the underlying clean images 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 deconvolution 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 similarity measure, which can be viewed as a Mahalanobis distance [11], to compute the distance between pairs of cryo-EM images. Second, we use the proposed Mahalanobis distance to improve the class averaging algorithm described in [9]. 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 a synthetic dataset at various noise levels and observe an improvement in the number of nearest neighbors correctly detected.

Thanks to XYZ agency for funding.

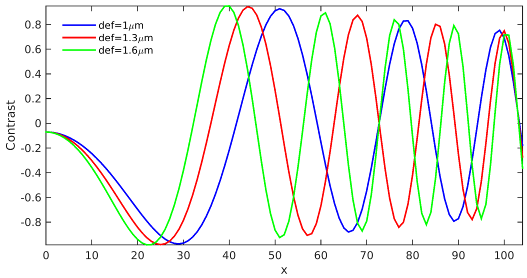

Fig. 1: CTF's for different values of the defocus.

2. BACKGROUND

2.1. Image Formation Model

Under the linear, weak phase approximation (see Chapt. 2 in [12]), the image formation model in cryo-EM is given by

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

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.(1) can be rewritten as

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

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 Chapt. 3 in [12])

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

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", and B is the B factor for the defocus dependent envelope function. CTF's corresponding to different defocus values have different zero crossings (see Fig.1). 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.

2.2. Rotationally Invariant Class Averaging

The procedure for class averaging, described in [9], consists of 3 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 [13]. 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[14] to improve the initial classification.

2.3. Covariance Wiener Filtering (CWF)

CWF was proposed in [10] as an algorithm to (i) estimate the CTF-corrected covariance matrix of the underlying clean 2D projection images (since phase flipping is not an optimal correction) and (ii) using the estimated covariance to solve the associated deconvolution problem in eqn. 2 to obtain denoised images, that are estimates of X_i for each i in eqn. 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

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

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

3. MAHALONOBIS DISTANCE

The Mahalanobis distance in statistics [11] 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 [15, 16].

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, say Y_1 and Y_2 in eqn.2. CTF correction is a challenging problem due to the numerous zero crossings of the CTF. 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 does not perform amplitude correction. In [10], a new approach for denoising and CTF correction in a single step was introduced, 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 suffer from incorrectly identified neighbors. One simple approach would be to simply use the Eucilidean distance between the CTF-corrected CWF coefficients of images obtained from denoising, as a measure of similarity. However, this measure does not take into account the structure of the covariance matrix of underlying clean data, that is, it does not account for the different covariances between different images in the dataset TODO re-word.

The main motivation of this paper is to introduce a Mahalanobis distance for cryo-EM images, as a similarity measure 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 [9].

In our statistical model, the underlying clean images $X_1, X_2, \dots X_n \in \mathbb{C}^d$ (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. We note that while the assumption of a Gaussian distribution does not hold in practice, it facilitates the derivation of this new measure. The justification of the new measure is its superiority over the older class averaging algorithm, as we demonstrate in Sec. 5.

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu, \Sigma) \\ N_i &\sim \mathcal{N}(0, \sigma^2 I_d) \end{aligned} \tag{5}$$

We denote the covariance of Y_i by K_i.

$$Cov(Y_i) = H_i \Sigma {H_i}^T + \sigma^2 I_n = K_i, \quad \text{for} \quad i = 1, \dots, n \tag{6}$$

Using the Guassian property, we have the following probability density functions (pdf) for $i = 1, \dots, n$

$$f_{X_i}(x_i) = P \exp\{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\}, \tag{7}$$

$$f_{N_i}(z_i) = Q \exp\{-\frac{1}{2}{z_i}^T \frac{1}{\sigma^2} z_i\}, \tag{8}$$

$$f_{Y_i}(y_i) = R_i \exp\{-\frac{1}{2}(y_i - H_i\mu)^T {K_1}^{-1}(y_i - H_i\mu)\}, \tag{9}$$

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

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} I & 0 \\ H_i & I \end{bmatrix} \times \begin{bmatrix} X_i \\ N_i \end{bmatrix} \tag{10}$$

$$\sim \mathcal{N}\left[\begin{bmatrix} \mu \\ H_i\mu \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma H_i^T \\ H_i\Sigma & H_i\Sigma H_i^T + \sigma^2 I \end{bmatrix}\right] \tag{11}$$

Using conditional distributions

$$f_{X_i|Y_i}(x_i|y_i) \sim \mathcal{N}(\alpha_i, L_i) \tag{12}$$

where

$$\begin{aligned} \alpha_i &= \mu + \Sigma H_i^T (H_i \Sigma H_i^T + \sigma^2 I)^{-1}(y_i - H_i\mu) \\ L_i &= \Sigma - \Sigma H_i^T (H_i \Sigma H_i^T + \sigma^2 I)^{-1} H_i \Sigma. \end{aligned} \tag{13}$$

So

$$x_i - x_j | y_i, y_j \sim \mathcal{N}(\alpha_i - \alpha_j, L_i + L_j) \tag{14}$$

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

$$\begin{aligned} &\Pr(||x_{ij}||_p < \epsilon | y_i, y_j) = \Pr(||x_{ij}||_p < \epsilon | y_i, y_j) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}|L_i + L_j|^{\frac{1}{2}}} \times \\ &\int_{B_p(0,\epsilon)} \exp\{-\frac{1}{2}(x_3 - \alpha_{ij})^T (L + M)^{-1}(x_3 - \alpha_{ij}\} dx_3 \end{aligned} \tag{15}$$

$$= \frac{\epsilon^d \text{Vol}(B_p(0,1))}{(2\pi)^{\frac{n}{2}}|L + M|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\alpha_{ij}^T (L + M)^{-1}\alpha_{ij}\} \tag{16}$$

where $B_p(0, \epsilon)$ is the ℓ_p ball of radius ϵ in $\mathbb{R}^d$ centered at the origin. 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. So we can define our similarity measure after taking the logarithm on both sides of eqn.(16), dropping out the constant term, and substituting back α_{ij}:

$$-\frac{1}{2}\log(|L + M|) - \frac{1}{2}(\alpha - \beta)^T (L + M)^{-1}(\alpha - \beta) \tag{17}$$

Notice the resemblance of the second term in eqn. 17 to the classical Mahalanobis distance [11]. This term takes into account the anisotropic nature of the covariance matrix by appropriately normalizing/scaling 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 Mahalanobis distance. Upto the first term, the similarity measure defined here is closely related to the one in [17].

4. ALGORITHM FOR IMPROVED CLASS AVERAGING USING MAHALANOBIS DISTANCE

We propose an improved class averaging algorithm that incorporates the Mahalanobis distance. 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 [9]. 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.

Add images for simulation

Algorithm 1 Improved Class Averaging

Require: A list of S nearest neighbor suspects for each image using initial classification [9].

procedure CLASSIFICATION USING MAHALONOBIS DISTANCE

  Compute the quantities α_i, L_i using estimates from Covariance Wiener Filtering (CWF) [10]

  For each image and its S aligned nearest neighbors, compute the Mahalanobis distance between the image and neighbors

  Rank S suspects according to the Mahalanobis distance, and choose the top K as nearest neighbors

procedure (OPTIONAL) IMPROVE NEAREST NEIGHBOR CLASSIFICATION USING VECTOR DIFFUSION MAPS (VDM) [14]

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 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. 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 number of true 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 by repeating this experiment at four different SNR's.

6. DISCUSSION

In this paper, we introduced a new similarity measure to compare CTF-affected cryo-EM images belonging to different de-

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

	VDM		No VDM	
SNR	This work	[9]	This work	[9]
1/60	34965	32113	34537	29219
1/100	17262	14431	16057	13706

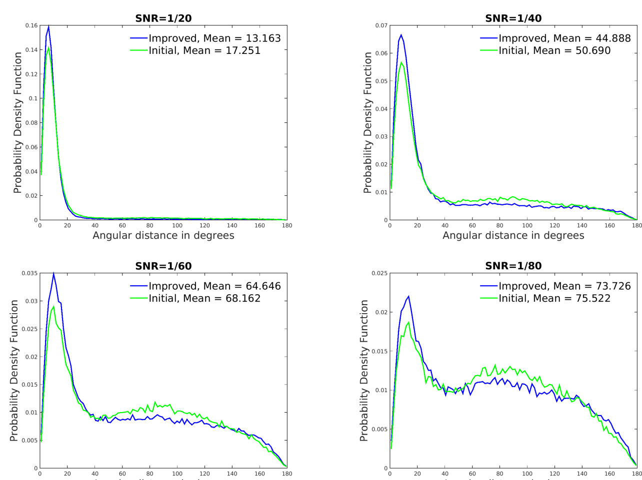

Fig. 2: The estimated probability density function of the angular distance (in degrees) between images classified into the same class by 1) Initial Classification and 2) Improved Classification using the Mahalanobis distance at different SNR's.

focus groups. TODO: mentioning that the resemblance of the affinity/similarity measure that was proposed in [20] and that in this work we provided a new probabilistic interpretation for it. The Mahalonobis distance proposed here can be used as a similarity measure for any manifold learning procedure [20, 17] such as diffusion maps [14, 21], with or without missing data.

7. REFERENCES

[1] 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.

[2] W. Kühlbrandt, "The resolution revolution," *Science*, vol. 343, no. 6178, pp. 1443–1444, 2014.

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

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

[5] P. A. Penczek, J. Zhu, and J. Frank, "A common-lines based method for determining orientations for $N > 3$ particle projections simultaneously," *Ultramicroscopy*, vol. 63, no. 3-4, pp. 205–218, 1996.

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

[7] M. van Heel and J. Frank, "Use of multivariate statistics in analysing the images of biological macromolecules," *Ultramicroscopy*, vol. 6, no. 2, pp. 187–194, 1981.

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

[9] 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.

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

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

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

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

[14] 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.

[15] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600 – 3612, 2008.

[16] X. Zhao, Y. Li, and Q. Zhao, "Mahalanobis distance based on fuzzy clustering algorithm for image segmentation," *Digit. Signal Process.*, vol. 43, no. C, pp. 8–16, Aug. 2015.

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

[18] 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.

[19] P.W. Jones, A. Osipov, and V. Rokhlin, "Randomized approximate nearest neighbors algorithm," *Proceedings of the National Academy of Sciences*, vol. 108, no. 38, pp. 15679–15686, 2011.

[20] R. Talmon and R.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.

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