Retrieving the parameters of cryo Electron Microscopy dataset in the heterogeneous ab-initio case

Yves Michels, Etienne Baudrier

RETRIEVING THE PARAMETERS OF CRYO ELECTRON MICROSCOPY DATASET IN THE HETEROGENEOUS AB-INITIO CASE

Yves Michels, Étienne Baudrier

ICube, University of Strasbourg, CNRS
300 Bd Sébastien Brant - CS 10413 - 67412 ILLKIRCH, FRANCE

ABSTRACT

A cryo Electron Microscopy dataset is composed on tomographic projections of an object (e.g. a macromolecule). The projection orientation information is unknown. The scope of this paper is the projection parameterization in the case of a deformable object. An overview of the parametrization methods is presented. Then a new approach based on manifold learning is detailed. Finally, an evaluation method for each substep of the parameterization is proposed. The resulting evaluation of the different methods on a 2D synthetic database shows the efficiency of our approach.

***Index Terms*—** tomography, cryo EM, heterogeneous set, parametrization, manifold learning

1. INTRODUCTION

The Cryo-Electron Microscopy (Cryo EM) is one of the main imaging modalities to study macromolecules. It enables, after reconstruction, to visualize the 3D volume of large macromolecules (1 000-10 000 atoms) with a resolution around the nanometer. The Cryo EM data consists of a large number (10 000-1 000 000) of images, each including a tomographic projection of a single specimen of the same macromolecule. For each images, the orientation and the state of deformation are unknown. Therefore, the reconstruction of a 4D object from the images obtained by cryo-EM is a difficult problem with important biological issues. If the case of a single object is well studied [1], the so-called heterogeneous case where the object has several states or a continuous set of states, is an active research field. In cryo EM, there are 3 groups of reconstruction methods: a) an ab initio analysis where the tomographic projections are sorted then a reconstruction by statistical methods [2], b) the tomographic projections are classified into subsets according to their orientation, then the heterogeneity analysis is made within each subset to identify different conformation [3], c) a posteriori analysis: a very large number of independent 3D maps are reconstructed after random selection of a small number of projections. Their comparison is used to locate areas of high variance in order to reclassify the projections according to these areas [4]. The b and c methods require a prior reconstruction. This prior reconstruction is used to refine projection parameters and then influences all the reconstructions. The frame of this paper is the ab-initio heterogeneous case, where all the projection parameters have to be estimated. In this case, the reconstruction process is generally composed of two steps: the estimation of the orientation and deformation parameters, then the object reconstruction from the projections and the parameter values. Only the parameter estimation step is studied in this paper.

The organization of the paper is as follows. Hereafter we describe briefly the state of the art for the parameter estimation. In Section 2, the dimension reduction step and the parameter estimation step of our estimation method are detailed. Then, in Section 3, experiments on both of these steps are conducted. A test on noise robustness is also made. A conclusion and perspectives end this paper in Section 4.

State of the art In the heterogeneous case without prior knowledge, the parameter estimation step relies on optimization cost [5], on likelihood optimization [6] or on a reduction of the dimension from the projection space to the the parameter space [7]. A study on the noise robustness of a dimension reduction based method is made in the 2D homogeneous case in [8]. In the case of a continuous deformation, the dimension reduction is interesting as the set of all the object conformations is a smooth manifold. Due to the direction parameter, the parameter space is not linear, which implies to choose a non-linear dimension reduction method.

Severals non-linear reduction dimension algorithms exist. This step plays an important role in the estimation accuracy, thus we pay a particular attention to the choice of the manifold learning method. Three general methods are adapted to our problem: i) The Graph Laplacian [9], widely used for manifold denoising [10] and dimension reduction, especially in the domain of tomography; ii) The methods based on the conservation of the local topology of the manifold. The chosen one for the comparison is the Hessian Locally Linear Embedding (H-LLE) [11]; iii) The methods based on the conservation of the distances in the manifold. The chosen one is Isomap [12].

2. PROPOSED METHOD

Let $\rho(p, x, y)$ be the density function of a planar deformable object, where $(x, y) \in \mathbb{R}^2$ are the planar coordinates and $p \in$

$[0,1]$ is a deformation parameter. Note that this model implies that we restrict ourselves to parameterizable deformation. In cryo EM, ρ stands for the electron absorption of the object. Its Radon transform $P_{(\theta,p)}(t)$ returns the value of the line integral of $\rho(p)$ along the parallel line L inclined at the angle θ with distance t from the origin [13].

$$P_{(\theta,p)}(t) = \int_{l \in L} \rho(p, x_L, y_L) \mathrm{d}l \tag{1}$$

For a given state of deformation p, tomographic reconstruction algorithms need the prior knowledge of the orientation θ for each projection. In this paper we consider n acquisitions $(\pi_i)_{1 \leqslant i \leqslant n}$ corresponding to unknown parameter couples (θ_i, p_i). We suppose here that these parameter couples are uniformly distributed in $[0, 2\pi[\times [0,1]$. Each acquisition is a vector in a m dimension space $\pi_i = (\pi_{i,j})_{1 \leqslant j \leqslant m}$. The set of projection acquisition is subject to high level of noise, modeled in our research by additive white Gaussian noise. In the case where the projections are parametrized with few parameters, the projections are lying in a low dimensional (LD) space whose the intrinsic dimensionality, d, is related to the number of freedom degrees. We call *representation*, the representation of a projection in a LD space.

2.1. Parameters estimation

In the 2D heterogeneous case, there are two unknown parameters with one in a modular space: $\theta \in \frac{\mathbb{R}}{2\pi\mathbb{Z}}$. As the estimation is done in the LD space, the ideal dimension reduction has to map the reduced projections on a cylinder. As there is a high noise level on real data, the represented manifold can have uncontrollable behavior and may not map on an easily parameterizable manifold. An illustration of this case is shown in Fig.1.

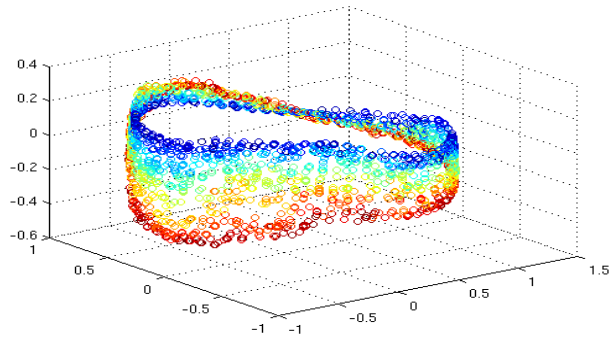


Fig. 1: Representation of 3200 noisily projections with SNR=10dB after a dimension reduction process.

To avoid this problem, we propose to parametrize the projection set in two steps. The first step parametrizes the angle and the next steps parametrize the deformation:

i) The angles are estimated with the same method as in [9], [14]. The orientations are estimated by ordering the representations in the first two dimensions where the representations are lying on a circle. We can define for each point an angle ϕ_l which respects:

$$X_l + iY_l = r(l)e^{i\phi_l} \tag{2}$$

where X_l and Y_l are the first two coordinates of the representation of the projection l. Within a constant, the order of the orientations θ_l is the same as the order of the angles ϕ_l.

ii) Once the orientations estimated, the set of projections is separated in several subsets corresponding to a segment of orientations $\hat{\theta} \in [\theta_{min}, \theta_{max}[$. A dimension reduction algorithm is then applied on each subset in order to obtain a 2-dimension representation of the projections where the deformation is easily estimable, Fig. 2.

iii) The representations are rotated in order to align the axes of equivalent angles along the first dimension as shown in Fig. 2. Then the slight remaining curvature is corrected by aligning the hulls along the second dimension. The hulls are modeled by a second order polynomial function. An illustration is given by the cross in magenta in Fig.2 Empirical experiences showed that our correction is not appropriate for subsets containing more than $n/6$ projections because of the complexity of the geometry of the manifold. However, the variance of the estimation is proportional of the inverse of the subset size. A reasonable trade off is to divide the dataset in 8 subsets of projections.

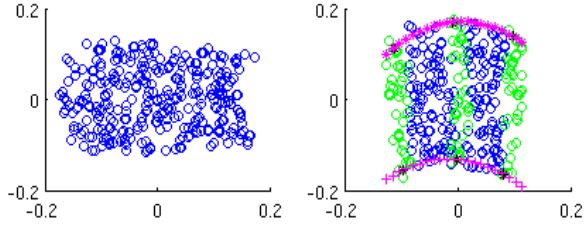


Fig. 2: (Left) Representation of a subset of 400 projections extracted from 3200 projections of our object with orientation and state of deformation uniformly distributed in $[0, 2\pi) \times [0,1]$. (Right) The same representation aligned by axes of equivalent angles. The two estimated hulls are represented with magenta crosses.

iv) The deformation parameters are estimated by ranking the representations along the first dimension for each subset. To use an unique definition of the deformation, the deformation estimations are aligned with the neighbored subsets. The alignment is illustrated in Fig.3

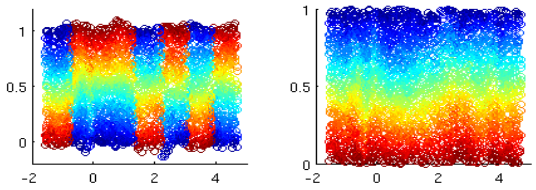


Fig. 3: Concatenation of 8 corrected representations of the projection subsets on the right and estimation of (θ, p) by ranking and registering the subsets. The color corresponds to the state of deformation.

2.2. Non-linear dimension reduction

The comparison of the manifold learning methods, presented in Section 3 gives Isomap as the best dimension reduction method to parametrize the projection set among the tested method. Then, in this section, our method is detailed using the Isomap method.

The use of *a priori* knowledge allows us to adapt the non-linear dimension reduction. In tomography, two projections with opposite orientations are symmetric:

$$P_{(\theta+\pi,p)}(t) = P_{(\theta,p)}(-t) \tag{3}$$

where $t = 0$ corresponds to the center of rotation of the object.

With a symmetrized Euclidean distance between the projections, it is possible to restrict the orientation parameter θ in $[0, \pi)$. The use of this distance decreases by 2 the standard deviation of the estimation[9].

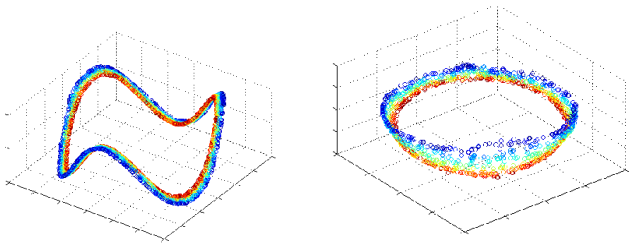
The dimension reduction relies on the conservation of the distances between the projections in the high dimensional space and the distances between the representations in the LD space. Nevertheless the distance used in the LD space is the Euclidean distance while the reference one is the geodesic distance. In general cases, the distance due to the deformation is smaller than the distance due to the orientation. As this property is conserved by the dimension reduction, the manifold of the whole dataset in the LD space is a thin cylinder. To improve the orientation estimation, the Euclidean distance can be estimated in the LD space with the formula:

$$D_{E_{l,h}} = \sqrt{2 - 2cos(G_{l,h})}, \tag{4}$$

with $(l, h) \in \{1, ..., n\}^2$ and where D_E is the estimated Euclidean distance in the LD space and G is the normalized geodesic distance estimated in the projections space, $max(G) = \pi$. The adaptations are summarized in Fig.4.

2.3. Noise robustness

In Cryo-EM, the dataset is subject to a high level of noise that is modeled by a white Gaussian noise with low Signal to Noise Ratio (SNR).



(a) Geodesic distance

(b) Estimated Euclidean distance

Fig. 4: Representations of 3200 projections issues from our object with orientation and state of deformation uniformly distributed in $[0, 2\pi) \times [0, 1]$. The color corresponds to the state of deformation.

The noise propagates to the distance graph where it may induce shortcuts that compromise the dimension-reduction process. The denoising occurs in 2 steps:

i) A denoising algorithm is used before our dimension reduction step. Because of the large number of projections, n, we choose a graph based denoising algorithm [10] which efficiently recovers the topology of the manifold and the respective position of each projections. Nevertheless, as this denoising method smooths the projections, the denoised projections are only used for the parameter estimation. The used denoising method depends on 2 parameters that are determined experimentally: the number of neighbors $k = 25$ used to compute the graph sets and a step to compute the Euler's scheme fixed to $\delta = 0.1$.

ii) All the neighbor which are not mutual are suppressed from the neighbor graph.

3. EXPERIMENTATION

This section proposes a comparison of the dimension reduction and an evaluation of our method. The dataset is composed of projections from ten 2-dimensional objects made using the $MolMov$ databank [15, 16, 17, 18] where the deformation is controlled by a parameter $p \in [0, 1]$ presented in Fig.5. To fit to the reality, the projections are taken at random orientations and deformations $(\theta_i, p_i)_{i\in[1,n]}$.

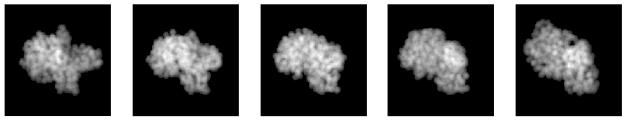


Fig. 5: One 2D object in 5 conformations corresponding to $p = \{0, 0.25, 0.5, 0.75, 1\}$.

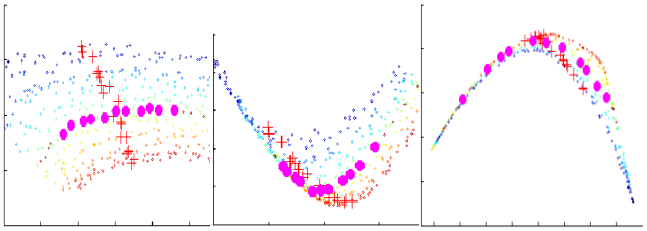
Comparison of the dimension-reduction methods The comparison is done on the ability to make, locally, the orienta-

	Isomap	H-LLE	Laplacian Graph
E_p	6.68	**8.12**	6.05
E_θ	**13.04**	3.39	3.10
C	0.217	**0.198**	0.480
ϵ	**0.0446**	0.0559	0.1580
ω	**16.05**	9.48	6.59

Table 1: Result of the evaluation of the dimension-reduction methods

tion and deformation parameter separable and estimable. Let θ_0 and p_0 be a given orientation and deformation state. Let define $S_p = \{R_{(\theta,p)}\ ,\ p \in [0,1], \theta \in [\theta_0 - d\theta, \theta_0 + d\theta]\}$ and $S_\theta = \{R_{(\theta,p_0)}\ ,\ \theta \in [\theta_0 - \alpha, \theta_0 + \alpha], p \in [p_0 - dp, p_0 + dp]\}$, where R is the LD representation of $P_{(\theta,p)}$, $d\theta > 0$ and $dp > 0$. We can define a distribution ellipse for S_p and S_θ. Lets a_θ and b_θ (respectively a_p and b_p) the first two principles axis of the distribution ellipse of S_θ (respectively S_p). The elongation of the point cloud is quantified by $E_\theta = |a_\theta|/|b_\theta|$, respectively $E_p = |a_p|/|b_p|$. The collinearity between the two points clouds is quantified by $C = | < a_p, a_\theta > |$ where $< .,. >$ is the scalar product. The ability of the dimension-reduction method to make the parameters estimable is evaluated by the error of the deformation parameter estimation, ϵ, made by ranking the representation of S_p along a_p. An illustration is given in Fig.6. the aim of the dimension-reduction step is to obtain high elongations E_θ and E_p, a low correlation C and a low error ϵ. Then we propose the following indicator ω to evaluate the dimension-reduction success.

$$\omega = \frac{(1-C)\sqrt{E_\theta E_p}}{0.5 - \epsilon}$$



(a) Isomap, $w = 14.8$ (b) H-LLE, $w = 5.7$ (c) Laplacian graph, $w = 2.7$

Fig. 6: Representation of the same subset of 380 projections by Isomap, H-LLE and Laplacian Graph methods. The points in S_p are marked by red crosses and the points in S_θ are marked by magenta circles.

The evaluation is done on 150 subsets of 380 projections from a dataset of 3000 projections. For each subset, a reduction dimension is applied and the features are calculated for $p_0 = 0.5$ and θ_0 which is the median of the projection orientations in the subset.

Noise robustness We apply the above algorithm to the estimation of the parameters $(\theta_i, p_i)_{i\in[1,n]}$ from the dataset. The evaluation of our method is done on the accuracy of the estimation measured by the mean of the absolute error of the estimation. Two factors can impact the quality of the estimation: the number of projections and the level of noise. Fig.7 represents the error made on a set of 4096 projections with different noise levels. For a given SNR, the noise is a Gaussian with zero mean and a variance, σ^2, that satisfies $SNR[dB] = 10\log(var_{signal}/\sigma^2)$, where var_{signal} is the variance of the signal without the edges $P_{(\theta_i,p_i)}(t_j) = 0$.

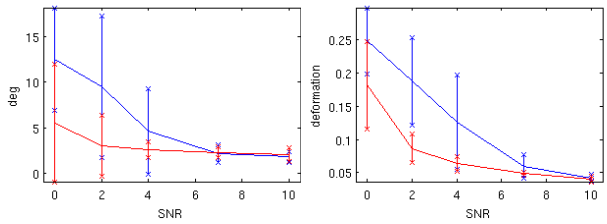


Fig. 7: Mean absolute error of the estimations in function of the SNR, with the denoising step, in red, without the denoising step, in blue. The curve is an average on 10 experiments, each corresponding to one 2D object, done on a set of 4096 projections $P_{(\theta,p)} \in \mathbb{R}^{217}$ taken at random (θ_i, p_i).

When the number of projections is too low, the estimation method suffers from artifacts on the estimation of p because of the size of the subsamples. For reasonable number of projections, $n > 2500$, the standard deviation of the error is around 2 degrees and 0.05 unity of deformation for a SNR higher than 7dB. For SNR lower than 4dB the estimation is highly impacted by the presence of non-detected shortcuts which requires to be processed.

4. CONCLUSION

This paper introduces a new method based on manifold learning to estimate the orientation and deformation states from a set of tomographic projections of a deformable object. The projection set is embedded in a low dimensional space where the projection orientation is estimated by first, then the deformation parameter is estimated on subsets of the projection set in the low dimensional space. An indicator is proposed to evaluate the ability of the method to separate the deformation from the orientation in the LD space. Our method is compared with state-of-the-art dimension-reduction methods on a 2-dimensional deformable object and the test shows that our method gives the best separation ability and accurate estimation even at high level of noise.

In future works, shortcut detection algorithms will be developed and the method will be extended to the 3D case and tested on real data.

5. REFERENCES

[1] J. Frank, *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. New York: Oxford Univ. Press, 2006.

[2] N. Elad, D. Clare, H. Saibil, and E. Orlova, "Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections," *J Struct Biol*, vol. 162, no. 1, pp. 108–120, 2008,

[3] M. Shatsky, R. Hall, E. Nogales, J. Malik, and S. Brenner, "Automated multi-model reconstruction from single-particle electron microscopy data," *J Struct Biol*, vol. 170, no. 1, pp. 98–108, 2010.

[4] B. Klaholz, A. Myasnikov, and M. Van Heel, "Visualization of release factor 3 on the ribosome during termination of protein synthesis," *Nature*, vol. 427, pp. 862–865, 2004.

[5] B. Ben Cheikh, E. Baudrier, and G. Frey, *Ab initio reconstruction of projection directions and volumes from heterogeneous data in cryo electron microscopy*, submitted.

[6] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, "Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization," *Nature Methods*, vol. 4, pp. 27–29, 2007.

[7] P. Schwander, R. Fung, and A. Ourmazd, "Conformations of macromolecules and their complexes from heterogeneous datasets," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, no. 1647, 2014.

[8] A. Singer and H. T. Wu, "Two-dimensional tomography from noisy projections taken at unknown random directions," *J Imaging Sci*, vol. 6, no. 1, pp. 136–175, Feb. 2013.

[9] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Trans Image Process*, vol. 17, no. 10, pp. 1891–1899, Oct. 2008.

[10] M. Hein and M. Maier, "Manifold Denoising," *Adv. Neural Inf. Process. Syst. 19*, pp. 561–568, 2007.

[11] D. L. Donoho and C. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high-dimensional data." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[12] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[13] A. Faridani, "Introduction to the Mathematics of Computed Tomography," *MSRI publications*, vol. 47, pp. 1–46, 2003.

[14] Y. Fang, M. Sun, S. V. N. Vishwanathan, and K. Ramani, "slle: Spherical locally linear embedding with applications to tomography," *Proc CVPR IEEE*, pp. 1129–1136, June 2011.

[15] "MolMovDB sets page," http://molmovdb.org/cgi-bin/movie.cgi?set=highMaxDev, accessed: 2016-05-09.

[16] W. G. Krebs and M. Gerstein, "The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework," *Polymer and Cell Dynamics*, vol. 28, no. 8, pp. 1665–1675, 2000.

[17] S. Flores, N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E. Z. Yu, M. Thorpe, and M. Gerstein, "The Database of Macromolecular Motions: new features added at the decade mark," *Nucleic Acids*, vol. 34, pp. 296–301, 2006.

[18] S. C. Flores and M. B. Gerstein, "Predicting protein ligand binding motions with the conformation explorer," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–12, 2011.