The generalized method of moments for multi-reference alignment

Asaf Abas, Tamir Bendory, and Nir Sharon

Abstract—This paper studies the application of the generalized method of moments (GMM) to multi-reference alignment (MRA): the problem of estimating a signal from its circularly-translated and noisy copies. We begin by proving that the GMM estimator maintains its asymptotic optimality for statistical models with group symmetry, including MRA. Then, we conduct a comprehensive numerical study and show that the GMM substantially outperforms the classical method of moments, whose application to MRA has been studied thoroughly in the literature. We also formulate the GMM to estimate a three-dimensional molecular structure using cryo-electron microscopy and present numerical results on simulated data.

Index Terms—generalized method of moments, multi-reference alignment, single-particle cryo-electron microscopy, orbit recovery problem

I. INTRODUCTION

Multi-reference alignment (MRA) is the problem of recovering a signal $x \in \mathbb{R}^L$ from

$$y_i = R_{s_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

where R_s circularly-translates the signal by s elements, i.e., $(R_s x)[j] = x[(j - s) \bmod L]$, and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$. The translations s_i are sampled from an unknown distribution ρ. Since the translations $R_{s_1}, \ldots, R_{s_N}$ are unknown, we cannot distinguish between (1) and the set of observations $y_i = R_{s_i - \tilde{s}}(R_{\tilde{s}} x) + \varepsilon_i$ for any $\tilde{s} \in \mathbb{Z}$. Thus, our goal is to estimate x, up to a circular translation. In group theory terminology, the set of signals $\{R_s x\}_{s=0}^{L-1}$ is called the *orbit* of the signal under the group of circular translations.

The MRA model is motivated by applications in signal processing [1] and structural biology [2], [3], [4], [5]. In particular, it has been demonstrated as a useful mathematical abstraction for single-particle cryo-electron microscopy (cryo-EM), an emerging technology to elucidate the 3-D structure of biomolecules [6], [7], [8], [9], [10]. Cryo-EM is the chief motivation of this paper, and is the main focus of Section V.

If the signal-to-noise ratio (SNR) is sufficiently high, one can recover the signal x by estimating the circular translations $R_{s_1}, \ldots, R_{s_N}$ using a variety of synchronization algorithms [11], [12], [13], [14], align them (i.e., undo the circular translations), and average out the noise. However, low SNR

hinders reliable shift estimation [15], [16], and thus in this regime one must estimate the signal directly.

The *method of moments* (MoM) is a classical statistical inference technique, tracing back to 1894 [17]. The MoM estimator, which is described in detail in Section II-A, results in a set of parameters whose moments agree with the empirical moments of the observed data $\{y_i\}_{i=1}^N$. A single-pass through the observations is required to compute the empirical moments. This single-pass requirement stands in contrast to mainstream parameter estimation techniques, such as maximum likelihood estimation, which usually iterates over the data. Thus, the MoM is an attractive computational framework for massive data sets. In addition, it was shown that in the low SNR regime, when SNR $\to 0$, $N \to \infty$, and L is fixed, the MoM achieves the optimal sample complexity of MRA, namely, the minimal number of observations needed for recovery up to an arbitrary precision [18], [19], [20], [21], [22], [23]. Notwithstanding, finding parameters that approximate the observable moments, namely moment fitting, frequently requires solving a system of nonlinear polynomial equations, a challenging computational task for high-dimensional data.

A standard method for fitting the analytic and the empirical moments is by minimizing a least-squares (LS) objective, which often results in a sub-optimal solution. As a remedy, this paper studies the *generalized method of moments* (GMM), which suggests matching the moments by a weighted LS objective [24], while providing an explicit expression of the optimal set of weights. Choosing the optimal weights enjoys appealing statistical properties, as discussed in Section III.

This work studies the GMM and its application to the MRA problem. Applying the GMM to MRA raises two main challenges: the dimensionality of the problem, which is typically high, and the solution's symmetry (i.e., the solution is defined up to a symmetry). In particular, classical GMM theory shows that the GMM provides an optimal estimator if a single set of parameters fits the observable moments. Namely, the polynomial system of equations has a unique solution. Unfortunately, this is never the case for MRA as the solution is defined up to a circular translation. Filling a theoretical gap, this work proves that the GMM retains its favorable statistical properties even when the statistical model has an intrinsic group symmetry.

The paper is organized as follows. In Section II, we describe in detail the MoM and the GMM estimators. Next, Section III discusses the properties of the GMM estimator and our theoretical work to extend it for models with intrinsic group symmetry. Section IV provides a comprehensive numerical study. Our study demonstrates that the GMM outperforms the

MoM in a variety of noise models and levels. We also provide a heuristic to predict when the performance gap between the GMM and the MoM is expected to be significant. Ultimately, Section V formulates the GMM for the problem of recovering a 3-D molecular structure using cryo-EM and presents initial numerical results.

II. THE GENERALIZED METHOD OF MOMENTS

A. The method of moments

Before introducing the GMM framework, we begin by presenting the classical MoM. Suppose that a random variable $y \in \mathbb{R}^r$ is drawn from a distribution which can be characterized by a set of parameters $\theta_0 \in \Theta$, where Θ is a compact space. The goal is to recover the unknown parameters θ_0 from N samples $y_1, \ldots, y_N$. In the MoM, the underlying idea is to estimate θ_0 from the first k_M empirical moments of the observations. We calculate the empirical moments from the data by averaging over the moments of individual observations. In particular, the k-th empirical moment is defined as

$$\hat{M}_k = \frac{1}{N} \sum_{i=1}^{N} y_i^{\otimes k}, \qquad (2)$$

where $y_i^{\otimes k}$ is a tensor with r^k entries. Each entry is given by $y_i^{\otimes k}(n) = \prod_{j=1}^{k} y_i(n_j)$, where $n = (n_1, \ldots, n_k)$ such that $1 \leq n_i \leq r$ for $i = 1, \ldots, k$. By the law of large numbers, for a large N we have,

$$\hat{M}_k \approx M_k(\theta_0) := \mathbb{E}\left[y^{\otimes k}\right], \quad k = 1, \ldots, k_M,$$

where $\mathbb{E}[\cdot]$ denotes expectation.

The MoM consists of two stages. First, one computes the first k_M observable moments $\hat{M}_k$ for $k = 1, \ldots, k_M$ from the data. In this work, we usually use the first two moments, that is, $k_M = 2$. In the second stage, we wish to find a set of parameters θ so that $\hat{M}_k \approx M_k(\theta)$ for $k = 1, \ldots, k_M$; this occasionally entails solving a system of polynomial equations. When a closed-form solution is not available, it is common to minimize a LS objective function,

$$\hat{\theta}_N^{LS} = \arg\min_{\theta \in \Theta} \sum_{k=1}^{k_M} ||M_k(\theta) - \hat{M}_k||_{\mathrm{F}}^2, \qquad (3)$$

where $\hat{\theta}_N^{LS}$ denotes the LS estimator, and Θ is the parameter space.

We note that the MoM is computationally attractive only when the number of observations N is much larger than the signal's length L. This is true since the computational complexity of the MoM is proportional to $L^{k_M} N$, whereas the complexity of methods that maximize the likelihood function (such as expectation-maximization) usually scales as TLN, where T is the number of passes through the data (which tends to increase as the SNR decreases [25], [26]).

B. The GMM framework

In its most simplified form, the GMM generalizes (3) by replacing the LS objective with a weighted LS. In particular, a specific choice of weights guarantees favorable asymptotic

statistical properties, such as the minimal asymptotic variance of the estimation error. We introduce these properties in detail in Section III.

Let us define the *moment function*, $f(\theta, y) \colon \Theta \times \mathbb{R}^r \to \mathbb{R}^q$. The moment function needs to be chosen such that its expectation value is zero only at a single point $\theta = \theta_0$. Namely,

$$\mathbb{E}\left[f(\theta, y)\right] = 0 \quad \text{if and only if} \quad \theta = \theta_0. \qquad (4)$$

We refer to (4) as the *uniqueness of the parameter set* condition. Henceforth, we choose the moment function to be

$$f(\theta, y_i) = \left[M_1(\theta) - y_i; \ldots; M_{k_M}(\theta) - y_i^{\otimes k_M}\right]. \qquad (5)$$

For convenience, we treat each moment as a column vector and the right hand side of (5) as their concatenation. For example $M_2(\cdot) \in \mathbb{R}^{r^2}$, and $f(\cdot)$ is in $\mathbb{R}^{r+r^2+\ldots+r^{k_M}}$. While we choose f as in (5), any moment function can be chosen as long as it satisfies the uniqueness condition and a few additional regularity conditions (as introduced in Appendix A). This flexibility enables the GMM to be applied to a wide range of problems, such as subspace estimation [27].

The estimated sample moment function is the average of f over N observations:

$$g_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} f(\theta, y_i). \qquad (6)$$

The GMM estimator is defined as the minimizer of the weighted LS expression,

$$\hat{\theta}_N = \arg\min_{\theta \in \Theta} g_N(\theta)^T W_N g_N(\theta). \qquad (7)$$

Here, W_N is a fixed positive semi-definite (PSD) matrix. Note that the LS estimator (3) is a special case of (7) when f is chosen as in (5) and $W_N = I$.

III. LARGE SAMPLE PROPERTIES

Before presenting the statistical properties of the GMM, we fix notation. We denote by $\xrightarrow{p}$ and $\xrightarrow{d}$ convergence in probability and in distribution, respectively. Let

$$S := \lim_{N \to \infty} \mathrm{Cov}\left[\sqrt{N} g_N(\theta_0)\right], \qquad (8)$$

be the covariance matrix of the estimated sample moment function (6) at the ground truth θ_0. We denote by $\{W_N\}_{N=1}^{\infty}$ a sequence of PSD matrices which converges almost surely to a positive definite (PD) matrix W. The expectation of the Jacobian of the moment function at the ground truth θ_0 is denoted by $G_0 = \mathbb{E}\left[\partial f(\theta_0, y)/\partial \theta^T\right]$.

A. GMM with a unique set of parameters

The large sample properties of the GMM estimator, under the uniqueness of the parameter set (4), were derived in [24], and are presented in the following theorem.

Theorem III.1. *Under the uniqueness of the parameter set* (4), *and the regularity conditions A.1-A.8 of Appendix A, the GMM estimator satisfies:*

   A. (Consistency) $\hat{\theta}_N \xrightarrow{p} \theta_0$.

B. (Asymptotic normality)

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, MSM^T),$$

where $M = [G_0^T W G_0]^{-1} G_0^T W$.

C. (Optimal choice of a weighting matrix) *The minimum asymptotic variance of $\hat{\theta}_N$ is given by $(G_0^T S^{-1} G_0)^{-1}$ and is attained by $W = S^{-1}$.*

Theorem III.1 provides a matrix W that guarantees a minimal asymptotic variance of the estimator's error. In Appendix A, we present the regularity conditions of Theorem III.1 in detail. While most of these conditions hold for the MRA model, e.g., continuity of $f(\cdot)$, the intrinsic symmetry of the MRA model (1) violates the uniqueness condition (4). In the next section, we extend Theorem III.1 and prove that the optimality of the GMM remains true even if there is a unique orbit of signals that fits the moments, rather than a unique signal.

The covariance matrix S of (8), which plays a central role in Theorem III.1, is required to be a PD matrix, see Appendix A. Therefore, the moment function must be chosen so that S is full-rank; see for example [28]. In this work, we noticed that if we remove the repeating entries of f (that appear due to the inherent symmetries of the moments), the covariance is indeed full rank (although, in some cases, ill-conditioned).

It is important to note that in practice, the ground truth θ_0 is unknown a priori, so we cannot use the optimal weighting matrix. A common heuristic is to replace (7) with an iterative scheme called iterative GMM. However, for our specific choice of moment function f (5), the covariance of g_N depends solely on the observations $\{y_i\}_{i=1}^N$, and not on the parameter set θ, namely,

$$\text{Cov}[g_N(\theta)] = \text{Cov}\left[y_i; \ldots; y_i^{\otimes k_M}\right]. \quad (9)$$

Therefore, one can compute both the moment function and the covariance matrix S (8) in a single pass. The algorithm is detailed in Algorithm 1.

Algorithm 1: The GMM estimator

Output: The estimated parameters $\hat{\theta}_{\text{GMM}}$
Input: A set of observations $\{y_i\}_{i=1}^N$
  1) Compute the estimated sample moment function g_N, given in (6)
  2) Compute the covariance $S = \text{Cov}\left[y_i; \ldots; y_i^{\otimes k_M}\right]$ and its inverse $W = S^{-1}$
  3) Solve $\arg\min_{\theta \in \Theta} g_N(\theta)^T W g_N(\theta)$

B. GMM with a unique orbit

We are now ready to introduce the main theoretical contribution of this paper: extending Theorem III.1 to the case where there is an orbit of signals that agrees with the observable moments. This is the case, for example, in the MRA model (1) as well as in cryo-EM [9].

Let $\mathcal{A}$ be a group acting on a vector space Θ. We denote the group action by $a \circ \theta$. To extend Theorem III.1, we make the following two assumptions.

Assumption III.2. *Global identification up to an orbit:*

$$\forall a \in \mathcal{A}: \ \mathbb{E}[f(\theta, y)] = 0 \text{ if and only if } \theta = a \circ \theta_0.$$

Assumption III.3. *Symmetry of the moment function:*

$$\forall a \in \mathcal{A}, \forall \theta \in \Theta, \forall i \in \{1, ..., N\}: \ f(a \circ \theta, y_i) = f(\theta, y_i).$$

The next theorem shows that under the above assumptions, the large sample properties of Theorem III.1 remain true for problems with group symmetry. In particular, Assumptions III.2 and III.3 are sufficient to guarantee consistency, as shown in Theorem III.4; the rest of the properties are direct corollaries. The proofs of Theorem III.4 and Corollary III.5 are provided in Appendices B and C, respectively.

Theorem III.4 (Consistency)**.** *Under Assumptions III.2, III.3, and the regularity conditions A.1-A.4 of Appendix A, there exists a sequence $\{a_N\}_{N=1}^\infty \subset \mathcal{A}$, such that:*

$$a_N \circ \hat{\theta}_N \xrightarrow{p} \theta_0,$$

where $\hat{\theta}_N$ is the GMM estimator.

Corollary III.5. *Under Assumptions III.2, III.3, and the regularity conditions A.1-A.8 of Appendix A, there exists a sequence $\{a_N\}_{N=1}^\infty \subset \mathcal{A}$, such that:*

A. (Asymptotic normality)

$$\sqrt{N}(a_N \circ \hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, MSM^T),$$

where $M = [G_0^T W G_0]^{-1} G_0^T W$.

B. (Optimal choice of weighting matrix) *The minimum asymptotic variance of $a_N \circ \hat{\theta}_N$ is given by $(G_0^T S^{-1} G_0)^{-1}$ and is attained for $W = S^{-1}$.*

IV. APPLICATION OF GMM TO MRA

A. Moments

The first two moments of the MRA model (1) are given by [21]:

$$M_1(x, \rho) = x * \rho = C_x \rho = C_\rho x,$$
$$M_2(x, \rho) = C_x D_\rho C_x^T + \Sigma.$$

Here, $*$ denotes a convolution, C_x is a circulant matrix whose first column is x, D_ρ is a diagonal matrix whose diagonal consists of the entries of ρ, and Σ is the covariance matrix of the noise. These moments can be estimated from the observations by the empirical moments $\hat{M}_1$ and $\hat{M}_2$ as in (2). By the law of large numbers, if N is large enough then $\hat{M}_1 \approx M_1$ and $\hat{M}_2 \approx M_2$.

In [21], it was shown that the first two moments suffice to recover the orbit of x and ρ, for almost any non-uniform distribution ρ and if the discrete Fourier transform (DFT) of x is non-vanishing. Therefore, a natural candidate for the moment function is

$$f(x, \rho, y_i) = \begin{pmatrix} M_1(x, \rho) - y_i \\ M_2(x, \rho) - y_i y_i^T \end{pmatrix}. \quad (10)$$

Since the second moment is a symmetric matrix, we remove all recurrent entries (e.g., eliminate the left lower triangle of the second moment matrix). Empirically, this modification results in a full-rank covariance S. With a slight abuse of

notation, we continue using the notation of (10) after removing the recurrent entries. Thus, $f \in \mathbb{R}^{L+(L+1)L/2}$.

Before diving into the numerical results, we need to verify that the MRA model (1) and the moment function (10) satisfy Assumptions III.2 and III.3. First, we note that Assumption III.2 is satisfied since the first two moments determine the orbit of the signal uniquely (under the aforementioned conditions). Let us define a group element a_i acting on $\theta = [\rho; x]$ by $a_i \circ [\rho; x] = [R_{-i}\rho; R_i x]$. Since M_1 and M_2 are invariant under this group action, the moment function (10) is invariant as well, and thus Assumption III.3 holds.

B. Experimental setting

Due to the inherent symmetry of the MRA model, we define the relative error as:

$$\text{error}(x, \hat{x}) = \min_{0 \le s \le L-1} \frac{\|R_s \hat{x} - x\|_2}{\|x\|_2}, \qquad (11)$$

where $\hat{x}$ is the signal estimate. The SNR is computed as

$$\text{SNR} = \frac{\|x\|_2^2}{\text{Trace}(\Sigma)},$$

where Σ is the covariance matrix of the noise term in (1).

For each SNR value, we conducted 100 trials. In each trial, we sampled a signal of length $L = 15$, drawn from a normal distribution $\mathcal{N}(0, I)$, and then normalized it such that $\|x\|_2 = 1$. We use low-dimensional signals since, in this regime, the MoM has clear computational advantages over methods that maximize the likelihood function, see Section II-A. In addition, it was recently shown that for high-dimensional signals, in contrast to the low-dimensional case, the sample complexity of the MRA model is not governed by moments [23]. The distribution ρ was drawn uniformly as an element over the simplex. Uniform distribution is merely a point on the continuous simplex, and therefore the distribution is almost surely non-uniform. Using the sampled signal and distribution, we generated $N = 100,000$ observations according to (1).

In the experiments, we compare the GMM estimator (7) with the classical MoM, corresponding to GMM with $W = I$; we refer to the latter as the LS estimator. We implemented both estimators using the general purpose interior point solver of MATLAB. While the objective functions of both the classical MoM and the GMM are non-convex, empirically, it seems that all local minima are global. Since the scale of the relative error changes drastically for different SNR levels, we measure the ratio between the relative errors of the GMM and the LS estimators. In the figures presented in this section, the blue and red lines represent, respectively, the mean and median of this relative error ratio. For clarification, when the ratio is greater than 1, the LS's relative error is bigger than the relative error of the GMM estimator. Namely, the GMM outperforms the LS. The upper and lower limits of the boxes denote the 75th and 25th percentiles, respectively. The dots represent trials whose ratios lie below the first quartile or above the third quartile. Lastly, the dashed line represents the LS's mean relative error, which corresponds to the right vertical axis. We focus on the range of SNR levels corresponding to relative errors that are
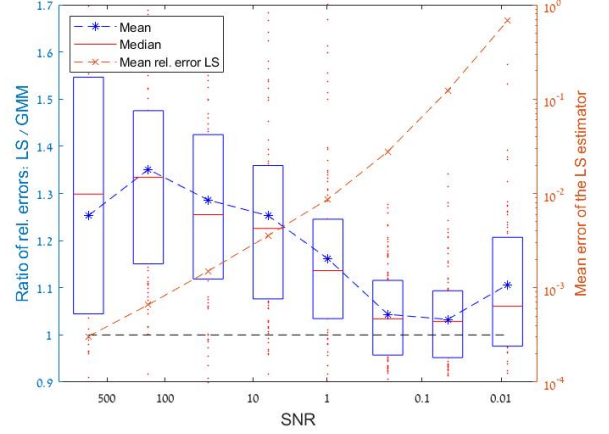

Fig. 1: **Homoscedastic noise model**: The ratio between the relative errors of the LS and the GMM estimators.

smaller than 1. The code to reproduce this section experiments is publicly available at https://github.com/abasasa/gmm-mra.

C. Homoscedastic noise

We start with a homoscedastic noise model where, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I)$; this noise model was considered by all previous works on MRA, e.g., [21], [26], [29].

The first experiment compares the performance of the LS and GMM estimators. The results are presented in Figure 1. As expected, the mean relative error of the estimators decreases as the SNR increases, as illustrated by the red dashed line. For high SNR levels, the GMM estimator outperforms the LS estimator by more than 20%. In low SNR levels, the GMM estimator has only a slight advantage over the LS estimator.

D. The effect of adding the third moment

This paper is mainly focused on using the first two moments: the minimal number of moments required for signal estimation when the distribution of translations is non-uniform [21]. This is also the number of moments we use for the cryo-EM experiments in Section V.

We now examine a natural question: what is the effect of adding the third moment to the moment function (10)? Figure 2 compares the performance of the GMM estimator with two and three moments. Adding the third moment indeed leads to significant improvement for all SNR levels, especially when the SNR is high. The superior numerical performance of adding the third moment to the moment function comes at a cost: the dimensionality (and thus the computational load) grows from $O(L^2)$ to $O(L^3)$. In particular, in our experiments, adding the third moment has increased the running time by a factor ranging from 1.5 to 5 (depending on the SNR). Figure 3 compares the performance of the GMM and LS estimators with the third moment. The trend of the ratio of relative errors is similar (besides a single point) to the same experiment with only the first two moments, as presented in Figure 1.

For the rest of the paper, we continue investigating the GMM estimator using only the first two moments.
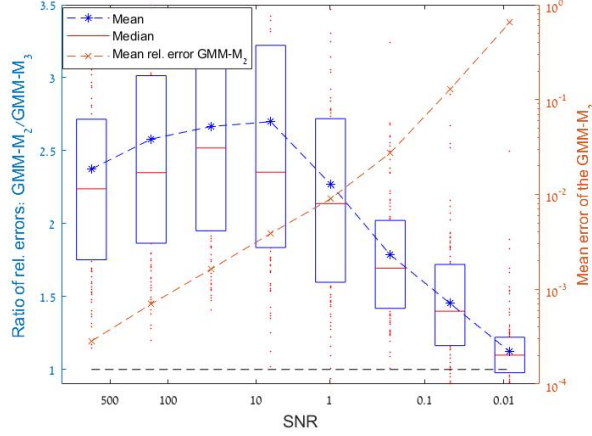
Fig. 2: **The effect of adding the third moment**: The ratio between the relative errors of the GMM estimator with the first three moments (GMM-M_3) and the GMM estimator with only the first two moments (GMM-M_2). Evidently, the GMM with three moments outperforms the GMM with two moments.
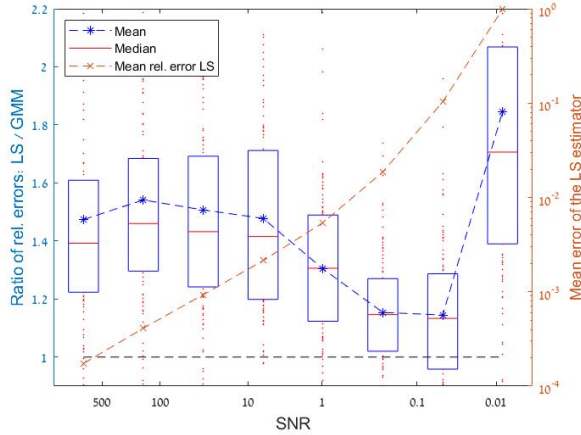

Fig. 3: **Homoscedastic noise model with the third moment**: The ratio of the relative errors of the LS and the GMM estimators when the moment function includes the third moment.

E. Heteroscedastic noise

Next, we investigate a scenario in which the noise term is distributed as $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_1)$, where Σ_1 is a diagonal matrix given by

$$
\Sigma_1 = \begin{bmatrix} \sigma^2 & & & \\ & 2\sigma^2 & & \\ & & \ddots & \\ & & & L\sigma^2 \end{bmatrix}. \tag{12}
$$

In this case, the noise level increases along the signal's entries. This model is similar to a popular noise model in cryo-EM [9]. Figure 4 compares the performance of the GMM and the LS methods and shows that the GMM outperforms the LS by at least 30% for most SNR levels. Note that the performance of the LS estimator is similar in both the homoscedastic and heteroscedastic noise models.
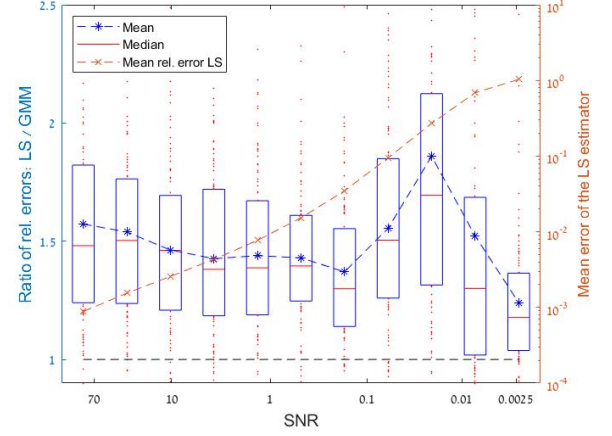

Fig. 4: **Heteroscedastic noise model**: The ratio between the relative errors of the LS and the GMM estimators.

F. When do we expect the GMM to outperform the MoM?

In the heteroscedastic noise model, the dispersion of eigenvalues is larger than in the homoscedastic case. Therefore, the effect of including the weighting matrix W_N in (7) leads to an increased performance gap between the LS and the GMM. Accordingly, we conjecture that in general, if W is "far" from the identity matrix in terms of large dispersion of eigenvalues, then the performance of the GMM will be significantly better compared to the LS estimator. To this end, we define the distance between a matrix $A \in \mathbb{R}^{n \times n}$ and the identity matrix by

$$
\delta(A) := \left[\sum_{j=1}^{n} \log^2 \left(\frac{\sqrt{n} \lambda_j(A)}{||A||_{\mathrm{F}}} \right) \right]^{\frac{1}{2}}, \tag{13}
$$

where $\lambda_j(A)$ is the j-th eigenvalue of A. This measure is based on the geodesic distance with respect to the Riemannian metric over the cone of PD matrices [30], with one additional normalization factor $\frac{\sqrt{n}}{||A||_{\mathrm{F}}}$. This factor assigns the same distance for matrices that are equal up to a scalar multiplication, i.e., $\delta(A) = \delta(cA)$ for any positive $c > 0$. In weighted LS, multiplication of the objective function by a scalar does not affect the estimator, and thus the modification is needed.

We compare the measure (13) of the weighting matrix (9), for the homoscedastic and heteroscedastic noise models; the results are presented in Figure 5. For the homoscedastic noise model, the minimum point of the geodesic distance is around SNR ~ 0.1. Indeed, Figure 1 shows that at this noise level, the performance of the GMM estimator performs similarly to the LS estimator. In addition, we observe that in high SNR homoscedastic noise and under heteroscedastic noise, large geodesic distance is positively correlated with superior performances of the GMM compared to LS. These observations support our hypothesis that the geodesic distance from the identity can be used as a heuristic tool to predict when it would be beneficial to use the GMM.
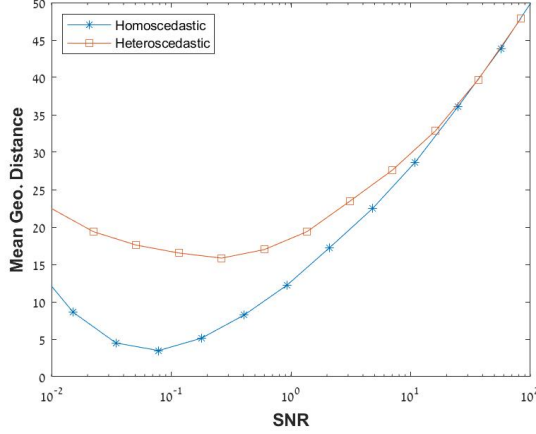
Fig. 5: Modified geodesic distance (13) of the weighting matrix W from the identity matrix as a function of the SNR. The two curves are of the homoscedastic and heteroscedastic noise models, averaged over 20 trials. In the homoscedastic noise model with SNR ≈ 0.1, where the geodesic distance between W and the identity matrix is small, the performance of the GMM and the LS estimators is similar.

G. MRA with outliers

Motivated by the abundance of outliers in cryo-EM datasets [9], we consider the following generative model

$$y_i = \begin{cases} R_{S_i}x + \epsilon_i & \text{w.p. } 1 - p_{\text{out}}, \\ \mathcal{N}(0, \Sigma_{\text{out}}) & \text{w.p. } p_{out}, \end{cases} \quad (14)$$

where p_{out} is the probability of an observation to be an outlier and Σ_{out} is the covariance matrix of the outliers' distribution.

Under this statistical model, the analytical moments read

$$M_1(x, \rho) = (1 - p_{\text{out}})C_\rho x,$$
$$M_2(x, \rho) = (1 - p_{\text{out}})C_x D_\rho C_x^T + (1 - p_{out})\Sigma + p_{\text{out}}\Sigma_{\text{out}}.$$

We assume that p_{out} and Σ_{out} are known. As before, for large N, the empirical moments approximate the analytical moments.

1) Weighted LAD optimization: Robustness to outliers is commonly obtained by replacing the LS objective with the weighted *least absolute deviations* (LAD) objective function [31], given by

$$\hat{\theta}_N^{LAD} = \arg\min_{\theta \in \Theta} |W_{\text{vec}}| \cdot |h(\theta, \{y_i\}_{i=1}^N)|. \quad (15)$$

Here, $|W_{\text{vec}}|$ is a fixed weighting vector, $\cdot$ is the scalar product, the absolute value is taken entrywise, and $h(\theta, \{y_i\}_{i=1}^N)$ is the loss function. Intuitively, one might suggest defining h of (15) as g_N from (6). However, it was shown in [32] that the estimation error of standard GMM based on ℓ_2 norm is lower than any other ℓ_p norm for a moment function that satisfies $\mathbb{E}[f(\theta_0, y_i)] = 0$[1]. Therefore, we take a different approach.

[1]This is not necessarily true for a biased moment function.

2) Geometric median estimator: We present an estimator based on the geometric median of the moments. The geometric median is a generalization of the univariate median to the multidimensional case, and is defined as

$$\widehat{GM}(\{y_i\}_{i=1}^N) = \arg\min_{z \in \mathbb{R}^q} \sum_{i=1}^N \left\| z - \begin{pmatrix} y_i \\ y_i y_i^T \end{pmatrix} \right\|_2, \quad (16)$$

where $q = L + (L+1)L/2$, the total number of elements of the first and second moment. Notice that as before, we treat $y_i y_i^T$ as a vector in $\mathbb{R}^{(L+1)L/2}$. The geometric median does not have a closed-form solution, and we approximate it using the Weiszfeld algorithm [33].

We estimate the empirical moments by (16), and use it in the loss function of the weighted LAD (15),

$$h_{GM}(x, \rho, \{y_i\}_{i=1}^N) = \begin{pmatrix} M_1(x, \rho) \\ M_2(x, \rho) \end{pmatrix} - \widehat{GM}(\{y_i\}_{i=1}^N). \quad (17)$$

We name this estimator the *geometric median estimator*. In order to construct W_{vec}, we use an alternating optimization scheme with two phases. The first phase minimizes (17) for a fixed W_{vec} (initialized by vectors of ones). Then, we update W_{vec} using the current estimation of x and ρ; see Algorithm 2. As mentioned in Section IV-F, a multiplication of W_{vec} by a scalar does not affect the GMM estimator, and therefore we minimize W_{vec} over the sphere in $\mathbb{R}^{L+(L+1)L/2}$, denoted here by $S^{L+(L+1)L/2-1}$, using Manopt's BFGS solver [34]. Empirically, updating W_{vec} once suffices.

Algorithm 2: Alternating weighted LAD

Output: The estimated parameter vector $\hat{\theta}$
Input:: A set of observations $\{y_i\}_{i=1}^N$, a loss function h, and number of steps n
  1) Initialize W_{vec} (a common choice is a vector of ones)
  2) Solve $\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}}|W_{\text{vec}}| \cdot |h(\theta, \{y_i\}_{i=1}^N)|$ (first iteration)
  3) **for** *j from 2 to n* **do**
    $\left|\begin{array}{l} W_{\text{vec}} = \underset{W \in S^{L+(L+1)L/2-1}}{\text{argmin}} |W| \cdot |h(\hat{\theta}, \{y_i\}_{i=1}^N)| \\ \hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}}|W_{\text{vec}}| \cdot |h(\theta, \{y_i\}_{i=1}^N)| \end{array}\right.$
    end

---

3) Numerical Results: The geometric median estimator is compared against the GMM in Figure 6 with $p_{\text{out}} = 0.2$ and $\Sigma_{\text{out}} = \frac{100}{L \cdot \text{SNR}}I$, where $L = 15$ is the signal length. As can be seen, for low SNR levels, the geometric median estimator exhibits superior numerical performance, while GMM works better in a high SNR environment. We observed similar results for different p_{out} values, however the SNR level of the transition point varies.

H. MRA with projection

The MRA with projection model is an extension of the standard MRA model (1), which includes an additional linear operator acting on the shifted signal. In this model, the i-th observation is given by

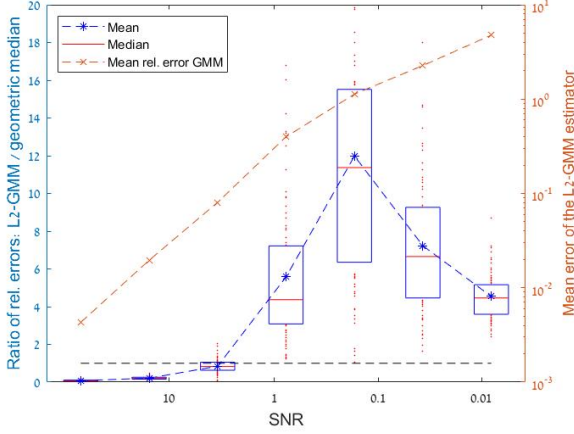$$y_i = PR_{s_i}x + \varepsilon_i, \quad (18)$$

Fig. 6: **Outliers noise model** (14): The ratio between the relative error of the GMM and the geometric median estimator for the outliers noise model (14) with $p_{\text{out}} = 0.2$.

where P is a fixed, known matrix of size $K \times L$. As in (1), the goal is to estimate $x \in \mathbb{R}^L$ from $y_1, \ldots, y_N \in \mathbb{R}^K$. The first and second moments of (18) are given by

$$M_1(x, \rho) = PC_\rho x,$$
$$M_2(x, \rho) = PC_x D_\rho C_x^T P^T + P\Sigma P^T.$$

Motivated by cryo-EM, we focus on a matrix P that samples only the first K entries of the shifted signal. That is, $PR_s x$ is a vector of length K consists of the entries $[(R_s x)_0, \ldots, (R_s x)_{K-1}]$.

Figure 7 compares the performance of the GMM and the LS estimators as a function of K (namely, how many entries are being kept) for a fixed noise level $\Sigma = 10^{-2} \cdot I_{K \times K}$. As can be seen, the GMM outperforms the LS by at least 20%.
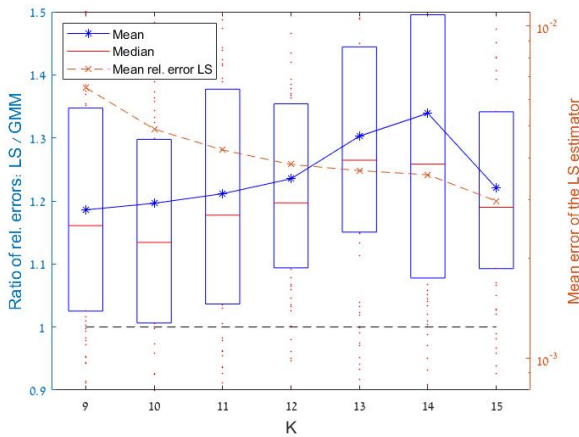

Fig. 7: **MRA with projection:** The ratio between the relative errors of the GMM and the LS estimators as a function of the measurement length K. The covariance of the noise is $\Sigma = 10^{-2} \cdot I_{K \times K}$.

V. GMM FOR CRYO-EM: A PROOF OF CONCEPT

As aforementioned, the main motivation of this work stems from the task of reconstructing the 3-D structure of molecular structures using single-particle cryo-EM. Building upon [35], we formulate the GMM for cryo-EM and show preliminary numerical results. We note that the goal of the MoM in the cryo-EM literature is not to reconstruct a high-resolution molecular structure, but only to quickly constitute a low-resolution, ab-initio model [9].

A. Mathematical background

Under some simplifying assumptions, the cryo-EM problem involves recovering a 3-D volume $\phi\colon \mathbb{R}^3 \to \mathbb{R}$ from a set of 2-D tomographic projection images. Each observation (projection image) is modeled as [6]

$$I_j = P(R_j^T \circ \phi) + \varepsilon_j, \tag{19}$$

where the projection operator $P\colon \mathbb{R}^3 \to \mathbb{R}^2$ is

$$P\phi(x_1, x_2) = \int_{-\infty}^{\infty} \phi(x_1, x_2, x_3)dx_3.$$

The term ε_j models additive noise term, and R_j is an element of the group of 3-D rotations $SO(3)$, which can be represented as a 3×3 orthogonal matrix acting by

$$[R_j^T \circ \phi(x_1, x_2, x_3)] = \phi(R_j[x_1 \ x_2 \ x_3]^T).$$

As in the MRA model (1), the group elements R_j are unknown. We assume that each image is sampled on an $n \times n$ Cartesian grid within the box $[-1, 1] \times [-1, 1]$.

The Fourier slice theorem states that the 3-D Fourier transform of a tomographic projection is equal to a slice of the Fourier transform of the volume. Thus, the Fourier transform of (19) reads

$$\widehat{I}_j(z_1, z_2) = (R_j^T \circ \widehat{\phi})(z_1, z_2, z_3)|_{z_3=0} + \widehat{\varepsilon}_j. \tag{20}$$

The common generative model of cryo-EM includes additional complications—such as the microscope's point spread function and heterogeneous mixture of molecules [9]—which are disregarded here for simplicity.

Following [35], we formulate the problem of recovering the 3-D structure ϕ from the first two moments. In polar coordinates, we can write the first empirical moment as

$$\hat{M}_1(r_1, \varphi_1) = \frac{1}{N} \sum_{j=1}^N \widehat{I}_j(r_1, \varphi_1),$$

and the second empirical moment

$$\hat{M}_2(r_1, \varphi_1, r_2, \varphi_2) =$$
$$\frac{1}{N} \sum_{j=1}^N \widehat{I}_j(r_1, \varphi_1, r_2, \varphi_2)\widehat{I}_j(r_1, \varphi_1, r_2, \varphi_2).$$

The first and second moment can be computed analytically

$$M_1(\phi, \rho) = \mathbb{E}\left[P(\widehat{R^T \circ \phi})\right],$$
$$M_2(\phi, \rho) = \mathbb{E}\left[P(\widehat{R^T \circ \phi}) \otimes P(\widehat{R^T \circ \phi})\right], \tag{21}$$

where the expectation is taken over the rotations and the noise. The explicit expressions of M_1 and M_2 of (21) are provided in Appendix D.
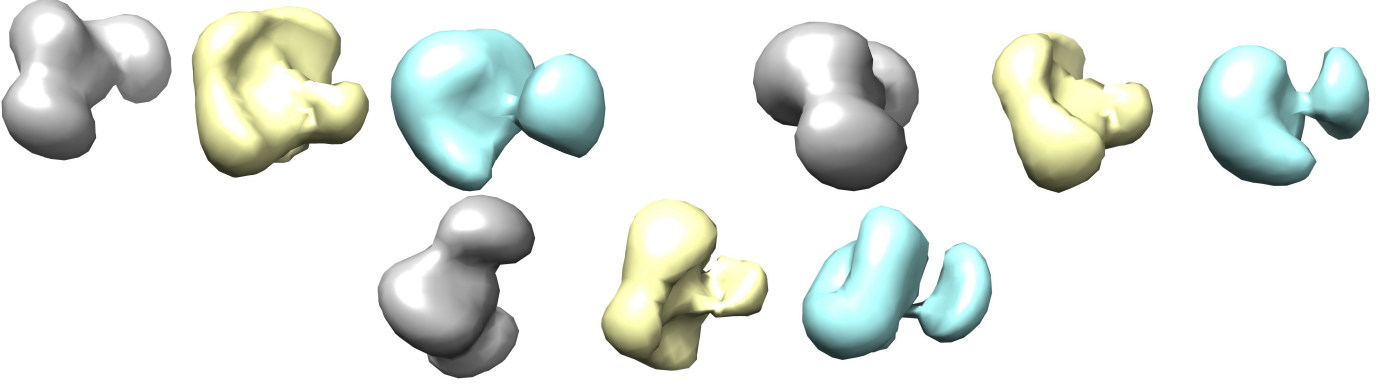
Fig. 8: Reconstructing a 3-D structure from noisy cryo-EM projections with SNR≈0.4. We present the structures from three different viewing directions. The ground truth, the LS estimation (MoM), and the GMM estimation appear in gray, yellow, and blue, respectively. The GMM outperforms the LS in both the FSC and relative error criteria. Its FSC value and relative error are 4.2 and 0.58, respectively. The FSC value of the LS estimator is 4.5, and its relative error is 0.61.

B. Numerical results

1) Evaluation metrics: We used two metrics: relative error and Fourier shell correlation (FSC). The relative error is computed similarly to (11), except that the alignment is computed over the group of 3-D rotations $SO(3)$. FSC is a common resolution measure in the cryo-EM field [36]. It measures cross-correlation coefficients between two 3-D volumes over corresponding shells in Fourier domain. Given two volumes, φ_1 and φ_2, the FSC in a shell κ is calculated using all voxels $\boldsymbol{\kappa}$ on this κ-th shell:

$$\text{FSC}(\kappa) = \frac{\sum_{||\boldsymbol{\kappa}||=\kappa} |\varphi_1(\boldsymbol{\kappa})||\overline{\varphi_2(\boldsymbol{\kappa})}|}{\sqrt{\sum_{||\boldsymbol{\kappa}||=\kappa} |\varphi_1(\boldsymbol{\kappa})|^2 \sum_{||\boldsymbol{\kappa}||=\kappa} |\varphi_2(\boldsymbol{\kappa})|^2}}.$$

Usually, FSC curves decrease with κ, and the resolution is determined as the point where the FSC curves drop below a prespecified value. In this work, we use a threshold of 0.5 [37]. Therefore, higher resolution is indicated by a smaller FSC value.

2) Example: Following [35], the LS estimator is formulated as

$$\arg\min_{\phi,\rho} \|M_1(\phi,\rho) - \hat{M}_1\|_F^2 + \lambda\|M_2(\phi,\rho) - \hat{M}_2\|_F^2,$$

where λ is a regularization parameter. As in the MRA problem, we define the moment function of the GMM estimator as

$$f(x,\rho,I_j) = \begin{pmatrix} M_1(\phi,\rho) - \hat{I}_j \\ M_2(\phi,\rho) - \hat{I}_j^{\otimes 2} + B \end{pmatrix}, \tag{22}$$

where B is an unbiasing term. We removed redundant entries, due to the inherent symmetries of the moments. In order to estimate the volumes, we used Manopt's trust-regions solver [34].

In the following experiment, the volume is a toy model, composed of five Gaussians over $\mathbb{R}^3$, whose high-frequencies were removed (see the blue volumes in Figure 8) [35]. The toy model is sampled on a $23 \times 23 \times 23$ Cartesian grid. The distribution of rotations ρ was drawn randomly such that ρ is invariant to in-plane rotations, i.e., ρ depends only on the viewing direction. We generated $N = 200,000$ observations

according to (19), and added an i.i.d. white Gaussian noise, corresponding to SNR ~ 0.4. Due to the simulated nature of the volume, the FSC's resolution units are measured in $1/$pixel's length.

The resulted volumes are depicted in Figure 8. The GMM outperforms the LS in both criteria. Its FSC value and relative error are 4.2 and 0.58, respectively. The FSC value of the LS estimator is 4.5, and its relative error is 0.61. The code is publicly available at https://github.com/ABASASA/GMM-Cryo.

C. Ill-conditioning of the weighting matrix

The condition number of the GMM's ideal weighting matrix W in the MRA model was around 10^4. Unfortunately, the condition number of W in the cryo-EM experiments was much higher. Figure 9 presents both the condition number of W and the geodesic distance between W and the identity matrix. As can be seen, as the SNR increases, so are the condition number and the distance to the identity matrix. In particular, for most SNR levels, W is ill-conditioned. This phenomenon must be considered when applying GMM to cryo-EM experimental datasets, as discussed in the next section.

VI. DISCUSSION

This paper is part of an ongoing effort to harness the favorable computational properties of the method of moments for constituting ab initio models of molecular structures using cryo-EM [5], [35], [38], [39], [40]. We use several MRA models as test cases and show that the GMM outperforms the classical method of moments. Also, we prove that the GMM retains its optimal statistical properties even when the signal is determined up to a group action, as in MRA and cryo-EM.

Our ultimate goal is to apply the GMM technique to constitute 3-D molecular structures using cryo-EM. However, our study reveals a major computational challenge: the moment function's covariance matrix is severely ill-conditioned. Thus, devising rigorous techniques to reduce the condition number is a future research direction toward applying the GMM framework to experimental cryo-EM datasets.
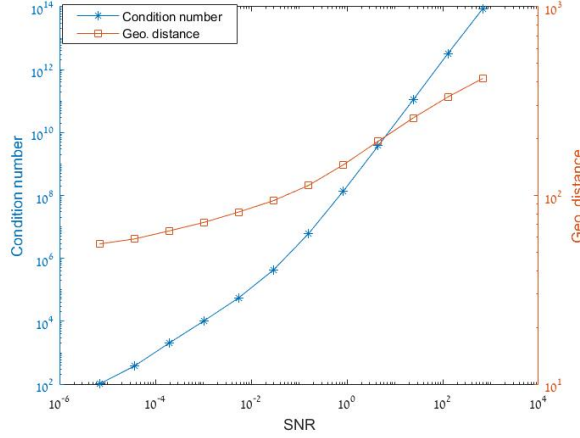
Fig. 9: The condition number and the geodesic distance (13), for the cryo-EM model (19), of the GMM weighting matrix (computed at the ground truth) for different levels of white noise.

In a broader perspective, we intend to study the GMM for additional signal processing tasks in which the method of moments plays an important role. Examples include super-resolution [41] and multi-target detection [16], [42], [43].

ACKNOWLEDGMENT

REFERENCES

[1] Joris Portegies Zwart, René van der Heiden, Sjoerd Gelsema, and Frans Groen. Fast translation invariant classification of HRR range profiles in a zero phase representation. *IEE Proceedings-Radar, Sonar and Navigation*, 150(6):411–418, 2003.

[2] Sjors HW Scheres, Mikel Valle, Rafael Nuñez, Carlos OS Sorzano, Roberto Marabini, Gabor T Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of molecular biology*, 348(1):139–149, 2005.

[3] Douglas L Theobald and Phillip A Steindel. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics*, 28(15):1972–1979, 2012.

[4] Chao Ma, Tamir Bendory, Nicolas Boumal, Fred Sigworth, and Amit Singer. Heterogeneous multireference alignment for images with application to 2D classification in single particle reconstruction. *IEEE Transactions on Image Processing*, 29:1699–1710, 2019.

[5] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer. Toward single particle reconstruction without particle picking: Breaking the detection limit. *arXiv preprint arXiv:1810.00226*, 2018.

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

[7] Eva Nogales and Sjors H.W. Scheres. Cryo-EM: A unique tool for the visualization of macromolecular complexity. *Molecular Cell*, 58(4):677–689, 2015.

[8] Kutti R. Vinothkumar and Richard Henderson. Single particle electron cryomicroscopy: trends, issues and future perspective. *Quarterly Reviews of Biophysics*, 49, 2016.

[9] Tamir Bendory, Alebrto Bartesaghi, and Amit Singer. Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Processing Magazine*, 37(2):58–76, 2020.

[10] Amit Singer and Fred J. Sigworth. Computational methods for single-particle electron cryomicroscopy. *Annual Review of Biomedical Data Science*, 3(1):163–190, 2020.

[11] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

[12] Afonso S Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 459–470, 2014.

[13] Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

[14] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

[15] Cecilia Aguerrebere, Mauricio Delbracio, Alberto Bartesaghi, and Guillermo Sapiro. Fundamental limits in multi-image alignment. *IEEE Transactions on Signal Processing*, 64(21):5707–5722, 2016.

[16] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer. Multi-target detection with application to cryo-electron microscopy. *Inverse Problems*, 35(10):104003, 2019.

[17] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[18] Afonso S Bandeira, Ben Blum-Smith, Joe Kileel, Amelia Perry, Jonathan Weed, and Alexander S Wein. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

[19] Afonso S Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. Optimal rates of estimation for multi-reference alignment. *Mathematical Statistics and Learning*, 2(1):25–75, 2020.

[20] Emmanuel Abbe, João M Pereira, and Amit Singer. Estimation in the group action channel. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 561–565. IEEE, 2018.

[21] Emmanuel Abbe, Tamir Bendory, William Leeb, João M. Pereira, Nir Sharon, and Amit Singer. Multireference alignment is easier with an aperiodic translation distribution. *IEEE Transactions on Information Theory*, 65(6):3565–3584, 2019.

[22] Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019.

[23] Elad Romanov, Tamir Bendory, and Or Ordentlich. Multi-reference alignment in high dimensions: sample complexity and phase transition. *SIAM Journal on Mathematics of Data Science*, 3(2):494–523, 2021.

[24] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029, 1982.

[25] Noam Janco and Tamir Bendory. An accelerated expectation-maximization for multi-reference alignment. *arXiv preprint arXiv:2105.07372*, 2021.

[26] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on Signal Processing*, 66(4):1037–1050, 2018.

[27] Jianqing Fan and Yiqiao Zhong. Optimal subspace estimation using overidentifying vectors via generalized method of moments. *arXiv preprint arXiv:1805.02826*, 2018.

[28] Howard E. Doran and Peter Schmidt. GMM estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. *Journal of Econometrics*, 133(1):387–409, 2006.

[29] Afonso S Bandeira, Yutong Chen, Roy R Lederman, and Amit Singer. Non-unique games over compact groups and orientation estimation in cryo-EM. *Inverse Problems*, 36(6):064002, 2020.

[30] Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

[31] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[32] Robert De Jong and Chirok Han. The properties of Lp-GMM estimators. *Econometric Theory*, pages 491–504, 2002.

[33] Endre Weiszfeld and Frank Plastria. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167(1):7–41, 2009.

[34] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014.

[35] Nir Sharon, Joe Kileel, Yuehaw Khoo, Boris Landa, and Amit Singer. Method of moments for 3D single particle ab initio modeling with non-uniform distribution of viewing angles. *Inverse Problems*, 36(4):044003, 2020.

[36] George Harauz and Marin van Heel. Exact filters for general geometry three dimensional reconstruction. *Optik (Stuttgart)*, 73(4):146–156, 1986.

[37] Marin van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of Structural Biology*, 151(3):250 – 262, 2005.

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

[39] Eitan Levin, Tamir Bendory, Nicolas Boumal, Joe Kileel, and Amit Singer. 3D ab initio modeling in cryo-EM by autocorrelation analysis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1569–1573. IEEE, 2018.

[40] Ti-Yen Lan, Nicolas Boumal, and Amit Singer. Random conical tilt reconstruction without particle picking in cryo-electron microscopy. *arXiv preprint arXiv:2101.03500*, 2021.

[41] Tamir Bendory, Ariel Jaffe, William Leeb, Nir Sharon, and Amit Singer. Super-resolution multi-reference alignment. *arXiv preprint arXiv:2006.15354*, 2020.

[42] Ti-Yen Lan, Tamir Bendory, Nicolas Boumal, and Amit Singer. Multi-target detection with an arbitrary spacing distribution. *IEEE Transactions on Signal Processing*, 68:1589–1601, 2020.

[43] Tamir Bendory, Ti-Yen Lan, Nicholas F Marshall, Iris Rukshin, and Amit Singer. Multi-target detection with rotations. *arXiv preprint arXiv:2101.07709*, 2021.

[44] Alastair R Hall. *Generalized method of moments*. Oxford university press, 2005.

[45] Larry C Andrews. *Special functions of mathematics for engineers*, volume 49. Spie Press, 1998.

[46] David Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty—IV: extensions to many dimensions; generalized prolate spheroidal functions. *Bell System Technical Journal*, 43(6):3009–3057, 1964.

[47] Gregory S Chirikjian and Alexander B Kyatkin. *Harmonic Analysis for Engineers and Applied Scientists: Updated and Expanded Edition*. Courier Dover Publications, 2016.

[48] Roy R Lederman. Numerical algorithms for the computation of generalized prolate spheroidal functions. *arXiv preprint arXiv:1710.02874*, 2017.

APPENDIX

A. Additional assumptions of the GMM

Note that by construction, g_N of (6) satisfies $\lim_{N\to\infty} g_N(\theta) = \mathbb{E}\left[f(\theta, y)\right] := g(\theta)$. In addition, we denote the corresponding quadratic forms,

$$Q(\theta) = g(\theta)^T W g(\theta), \quad Q_N(\theta) = g_N(\theta)^T W_N g_N(\theta),$$

and the Jacobians,

$$G(\theta) = \mathbb{E}\left[\partial f(\theta, y)/\partial \theta^T\right], \; G_N(\theta) = \frac{1}{N}\sum_{i=1}^N \partial f(\theta, y_i)/\partial \theta^T.$$

Notice that we also have $\lim_{N\to\infty} Q_N(\theta) = Q(\theta)$ and $\lim_{N\to\infty} G_N(\theta) = G(\theta)$.

The favorable statistical properties of the GMM rely on several conditions, including the uniqueness of a set of parameters (4), see Section III. For completeness, we now present the rest of the regularity conditions required for Theorem III.1 [24], [44]. For the first assumption, we say that a sequence of random variables with a joint cumulative distribution function F is strictly stationary if, for any τ, n, $t_1, \ldots, t_n \in \mathbb{N}$,

$$F(y_{t_1}, y_{t_2}, ..., y_{t_n}) = F(y_{t_1+\tau}, y_{t_2+\tau}, ..., y_{t_n+\tau}).$$

The above sequence (or process) of random variables is called ergodic if $\mathbb{E}\left[||\frac{1}{N}\sum_{i=1}^N y_i - \mathbb{E}\left[y_i\right]||^2\right] \to 0$, as $N \to \infty$.

Assumption A.1. *The sequence $\{y_i \mid i \in \mathbb{N}\}$ is strictly stationary and ergodic.*

Assumption A.2. *(i) Θ is a compact space and (ii) θ_0 is an interior point in Θ.*

Assumption A.3 (Regularity conditions for f)**.** *The moment function, $f: \Theta \times \mathbb{R}^r \to \mathbb{R}^q$, satisfies: (i) $f(\cdot, y)$ is continuous on Θ for every $y \in \mathbb{R}^r$; (ii) $g(\cdot)$ exists and finite for every $\theta \in \Theta$; (iii) $g(\cdot)$ is continuous on Θ and (iv) $\mathbb{E}\left[\sup_{\theta\in\Theta} \|f(\theta, y)\|\right]$ is finite.*

Assumption A.4 (Properties of the weighting matrix)**.** *$\{W_N\}_{N=1}^\infty$ is a sequence of PSD matrices, and W is a PD matrix such that $W_N \xrightarrow{p} W$.*

Assumption A.5 (Regularity conditions for $\partial f/\partial \theta^T$)**.** *(i) The Jacobian matrix $\partial f(\theta, y)/\partial \theta^T$ exists and continuous on Θ for every $y \in \mathbb{R}^r$ and (ii) $\mathbb{E}\left[\partial f(\theta, y)/\partial \theta^T\right]$ exists and finite.*

Assumption A.6 (Existence of the covariance)**.** *(i) $\mathbb{E}\left[f(\theta, y)f(\theta, y)^T\right]$ exists and finite; in addition (ii) S of (8) exists and it is a PD matrix.*

Assumption A.7 (Continuity of G)**.** *$G(\theta)$ is continuous on some neighborhood N_δ of θ_0 and its symmetries under the group $\mathcal{A}$, given by*

$$N_\delta(\theta_0) = \{\theta \in \Theta : \exists a \in \mathcal{A}, \|\theta - a \circ \theta_0\| < \delta\} =$$
$$= \cup_{a\in\mathcal{A}}\{\theta \in \Theta : \|\theta - a \circ \theta_0\| < \delta\}.$$

Assumption A.8 (Uniform Convergence of G_N)**.** *For N_δ of Assumption A.7, $G_N(\theta)$ satisfies*

$$\sup_{\theta\in N_\delta} \|G_N(\theta) - \mathbb{E}[f(\theta, y)/\partial \theta^T]\| \xrightarrow{p} 0.$$

B. Proof of Theorem III.4

In this section, we prove consistency for the GMM estimator under Assumptions III.2 and III.3. The proof is based on similar arguments as in [27] and [44].

Lemma A.9 (Uniform Convergence in Probability of $Q_N(\theta)$)**.** *Under Assumptions III.2, III.3 and Assumptions A.1, A.2, A.3 and A.4, we have*

$$\sup_{\theta\in\Theta} |Q_N(\theta) - Q(\theta)| \xrightarrow{p} 0. \tag{23}$$

Proof. Since $g(\cdot) = \mathbb{E}\left[f(\cdot, y)\right]$, we have

$$\|g_N(\theta) - \mathbb{E}\left[f(\theta, y)\right]\|_2 \to 0, \quad \theta \in \Theta.$$

By definition,

$$\sup_{\theta\in\Theta} |Q_N(\theta) - Q(\theta)| =$$
$$\sup_{\theta\in\Theta} |g_N^T W_N g_N - \mathbb{E}\left[f(\theta, y)\right]^T W \mathbb{E}\left[f(\theta, y)\right]|. \tag{24}$$

Since the terms of the right hand side of (24) are bounded and arbitrary small as $N \to \infty$, we obtain (23), as required. $\square$

Proof of consistency. Let $\delta > 0$ be a small enough constant such that Assumption A.7 holds. By definition, $N_\delta(\theta_0)$ is in the δ-neighborhood of θ_0 and its orbit. Then, it is an open set as an union of open sets under the standard metric in $\mathbb{R}^p$.

Therefore, its complementary set $[N_\delta(\theta_0)]^C$ is a closed set in Θ. Since $[N_\delta(\theta_0)]^C$ is a closed set in a compact set $\Theta \subset \mathbb{R}^p$, by Assumption A.2, $[N_\delta(\theta_0)]^C$ is also compact. According to Assumption A.3, the function g is continuous, and therefore $Q(\theta)$ is continuous as well. Using the extreme value theorem, $Q(\theta)$ has a minimum over a compact space $[N_\delta(\theta_0)]^C$.

Let us define $\varepsilon := \inf_{\theta \notin N_\delta(\theta_0)} Q(\theta)$, the infimum of the asymptotic objective function outside the neighborhood. First, we prove that $\varepsilon > 0$. By Assumption A.4, W is a constant PD matrix. Then, $Q(\theta) = 0 \iff g(\theta) = 0$. However, by Assumption III.2, $g(\theta) = 0$ only for a θ in the orbit of θ_0, hence $\theta \in N_\delta(\theta_0)$. Therefore, $\varepsilon > 0$.

By Lemma A.9, $\sup_{\theta \in \Theta} |Q_N(\theta) - Q(\theta)| \xrightarrow{p} 0$, so we can choose a large enough N such that: $|Q_N(\theta) - Q(\theta)| \leq \frac{\varepsilon}{3}$. In addition, by definition $Q_N(\hat{\theta}_N) \leq Q_N(\theta_0)$, since $\hat{\theta}_N$ is the minimum of Q_N. Then, for large enough N we get:

$$Q(\hat{\theta}_N) \leq Q_N(\hat{\theta}_N) + \frac{\varepsilon}{3} \leq Q_N(\theta_0) + \frac{\varepsilon}{3} \leq Q(\theta_0) + \frac{2\varepsilon}{3} < \varepsilon.$$

In other words, $\hat{\theta}_N \in N_\delta(\theta_0)$. Since δ was chosen arbitrarily, we conclude that there exists a sequence $\{a_N\}_{N=1}^\infty \subset \mathcal{A}$ such that $a_N \circ \hat{\theta}_N \xrightarrow{p} \theta_0$. $\quad\square$

C. Proof of Corollary III.5

The following proofs extend the proofs of [44] to uniqueness of an orbit solution. It is done by replacing $\hat{\theta}_N$ by $a_N \circ \hat{\theta}_N$.

Proof of asymptotic normality of the parameters estimator.
By Assumption III.3, we have $G_N(a_N \circ \hat{\theta}_N) = G_N(\hat{\theta}_N)$. From Theorem III.4 and the mean value theorem, there exists $a_N \in \mathcal{A}$ such that:

$$g_N(a_N \circ \hat{\theta}_N) = g_N(\theta_0) \\ + G_N(a_N \circ \hat{\theta}_N^\lambda)(a_N \circ \hat{\theta}_N - \theta_0). \quad (25)$$

Here, $\hat{\theta}_N^\lambda$ exists according to the mean value theorem. Note that the i-th row of $G_N(a_N \circ \hat{\theta}_N^\lambda)$ is the corresponding row to $G_N(\beta^{(i)})$, where $\beta^{(i)} = \lambda_{N,i}\theta_0^{(i)} + (1 - \lambda_{N,i})[a_N \circ \hat{\theta}_N]^{(i)}$. Multiplying (25) by $G_N(\hat{\theta}_N)^T W_N = G_N(a_N \circ \hat{\theta}_N)^T W_N$ yields

$$G_N(\hat{\theta}_N)^T W_N g_N(\hat{\theta}_N) = G_N(a_N \circ \hat{\theta}_N)^T W_N g_N(\theta_0) \\ + G_N(a_N \circ \hat{\theta}_N)^T W_N G_N(a_N \circ \hat{\theta}_N^\lambda)(a_N \circ \hat{\theta}_N - \theta_0).$$

By definition, $\hat{\theta}_N$ is the minimum of $Q_N(\cdot)$, and therefore $G_N(\hat{\theta}_N)^T W_N g_N(\hat{\theta}_N) = 0$. Thus,

$$\sqrt{N}(a_N \circ \hat{\theta}_N - \theta_0) = -[G_N(a_N \circ \hat{\theta}_N)^T W_N \\ G_N(a_N \circ \hat{\theta}_N^\lambda))]^{-1} G_N(a_N \circ \hat{\theta}_N)^T W_N \sqrt{N} g_N(\theta_0). \quad (26)$$

From the consistency of the estimator, we obtain:

$$G_N(a_N \circ \hat{\theta}_N) \xrightarrow{p} G_0,$$
$$G_N(a_N \circ \hat{\theta}_N^\lambda) \xrightarrow{p} G_0.$$

Next, we denote

$$M_N := [G_N(a_N \circ \hat{\theta}_N)^T W_N G_N(a_N \circ \hat{\theta}_N^\lambda))]^{-1} \\ \times G_N(a_N \circ \hat{\theta}_N)^T W_N \sqrt{N},$$

hence

$$\lim_{N \to \infty} M_N = [G_0^T W G_0]^{-1} G_0^T W.$$

The right hand side of (26) can be expressed as $-M_N g_N(\theta_0)$. Finally, from Slutsky's theorem and the central limit theorem, we derive the desired results. $\quad\square$

Proof of optimal choice of the weighting matrix. Let $\hat{\theta}_N(W)$ be the GMM estimator with the weighting matrix W. We denote by $V(W)$ the variance of the limiting distribution of $\sqrt{N}[a_N \circ \hat{\theta}_N(W) - \theta_0]$. We prove that $V(S^{-1})$ is the minimum asymptotic variance, which is equivalent to prove that $V(W) - V(S^{-1})$ is a PSD matrix, for any PD matrix W.

According to Theorem III.4, we have

$$\sqrt{N}[a_N \circ \hat{\theta}_N(W) - \theta_0] \to V(W),$$
$$\sqrt{N}[b_N \circ \hat{\theta}_N(S^{-1}) - \theta_0] \to V(S^{-1}),$$

when $N \to \infty$. Here, $\{a_N\}_{N=1}^\infty$ and $\{b_N\}_{N=1}^\infty$ are sequences in $\mathcal{A}$ such that $a_N \circ \hat{\theta}_N(W) \xrightarrow{p} \theta_0$ and $b_N \circ \hat{\theta}_N(S^{-1}) \xrightarrow{p} \theta_0$. We start by relating $\sqrt{N}[a_N \circ \hat{\theta}_N(W) - \theta_0]$ and $\sqrt{N}[b_N \circ \hat{\theta}_N(S^{-1}) - \theta_0]$ by

$$\sqrt{N}[a_N \circ \hat{\theta}_N(W) - \theta_0] = \sqrt{N}[b_N \circ \hat{\theta}_N(S^{-1}) - \theta_0] \\ + \sqrt{N}[a_N \circ \hat{\theta}_N(W) - b_N \circ \hat{\theta}_N(S^{-1})]. \quad (27)$$

From the proof of Corollary III.5-A,

$$\sqrt{N}[a_N \circ \hat{\theta}_N(W) - \theta_0] = -\sqrt{N} M(W) g_N(\theta_0) + o_p(1),$$

where $M(W) = [G_0^T W G_0]^{-1} G_0^T W$ and $o_p(1)$ represents a sequence of vectors $\{a_N\}_{N=1}^\infty \subset \mathbb{R}^{p \times 1}$ such that for each entry $\lim_{N\to\infty}(a_N)_i = 0$. Therefore,

$$\sqrt{N}[a_N \circ \hat{\theta}_N(W) - b_N \circ \hat{\theta}_N(S^{-1})] = \\ -\sqrt{N}[M(W) - M(S^{-1})]g_N(\theta_0) + o_p(1).$$

Now, we apply the asymptotic covariance of (27) and get:

$$V(W) - V(S^{-1}) = V_1 + C + C^T,$$

where $V_1 = \lim_{N\to\infty} \text{Var}\left[[M(W) - M(S^{-1})]\sqrt{N}g_N(\theta_0)\right]$, and

$$C = \lim_{N \to \infty} \text{Cov}[\sqrt{N}[M(W) - M(S^{-1})]g_N(\theta_0), \\ \sqrt{N}M(S^{-1})g_N(\theta_0)].$$

By construction, V_1 is a PSD matrix. For the choice of $W = S^{-1}$, it is easy to see that $C = 0$. Thus, we get the desired result. $\quad\square$

D. Computational framework for cryo-EM

We follow the framework developed in [35]. This section provides the complementary details which were omitted from Section V.

1) Basis for the volume $\widehat{\phi}$: We represent the Fourier transform of ϕ by

$$\widehat{\phi}(\kappa,\theta,\varphi) = \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell}\sum_{s=1}^{S(\ell)} A_{\ell,m,s}F_{\ell,s}(\kappa)Y_{\ell}^{m}(\theta,\varphi), \quad (28)$$

where $Y_{\ell}^{m}(\theta,\varphi)$ are the complex spherical harmonics, L is the volume's bandlimit, and $A_{\ell,m,s}$ are the expansion coefficients. The radial frequency functions $F_{\ell,s}$ are orthogonal for each fixed ℓ, where $s = 1,\ldots,S(\ell)$ is referred to as the radial index. Popular choices of radial functions include the spherical Bessel functions [45], which are eigenfunctions of the Laplacian on a closed ball with Dirichlet boundary conditions, as well as the radial components of 3-D prolate spheroidal wave functions [46].

We assume that the volume is band-limited with Fourier coefficients supported within the radius of size $\frac{\pi n}{2}$, where n is the size of the Cartesian grid of the projection images I_j, $n \times n$. Note that since ϕ is real valued, its Fourier transform is conjugate-symmetric, which imposes restrictions on the coefficients $A_{\ell,m,s}$.

The advantage of expanding $\widehat{\phi}$ in terms of spherical harmonics is that the space of degree ℓ spherical harmonics is closed under rotation. In particular, rotating a spherical harmonic by $R \in SO(3)$ can be expressed as

$$R^T \cdot Y_{\ell}^m(x) = Y_{\ell}^m(Rx)$$
$$= \sum_{m'=-\ell}^{\ell} U_{m,m'}^{\ell}(R)Y_{\ell}^{m'}(x), \ x \in S^2, \quad (29)$$

where $U^{\ell}(R) \in \mathbb{C}^{(2\ell+1)\times(2\ell+1)}$ are the Wigner matrices (see [47, p. 343]).

2) Basis for the probability distribution of rotations: We assume the probability density ρ over $SO(3)$ is a smooth, band-limited function, which can be expressed as

$$\rho(R) = \sum_{p=0}^{P}\sum_{u,v=-p}^{p} B_{p,u,v}U_{u,v}^p(R), \ R \in SO(3). \quad (30)$$

By the Peter-Weyl theorem, $\{U^{\ell}(R)\}_{\ell=0}^{L}$ form an orthonormal basis of $L^2(SO(3))$. The cutoff P is the band limit of ρ. Following the arguments from [35], we assume that $P \leq 2L$.

3) Basis for the 2-D images $\widehat{I}_j$: Next, we represent $\widehat{I}_j$ using a function space which is closed under in-plane rotations, represented as elements of $SO(2)$. By the Peter-Weyl theorem, we can can expand a band-limited image $\widehat{I}_j$:

$$\widehat{I}_j(\kappa,\varphi) = \sum_{q=-Q}^{Q}\sum_{t=1}^{T(q)} a_{q,t}^j f_{q,t}(\kappa)e^{iq\varphi}. \quad (31)$$

Here, the radial frequency functions $f_{q,t}$, for fixed q, are taken to be an orthonormal basis. Specifically, we choose $f_{q,t}$ to be the radial components of the 2-D prolate spheroidal wave functions [46]. Following [35], we take $Q = L$.

4) Representation of the moments: We present the connection between the first two moments of the observed images and the coefficients $\{A_{\ell,m,s}\}_{\ell,m,s}$ and $\{B_p,u,v\}_{p,u,v}$ of the volume and the distribution of rotations, respectively.

We index the images in terms of $R \in SO(3)$ (instead of j in (19)):

$$\widehat{I}_R(\kappa,\varphi) = \sum_{q=-Q}^{Q}\sum_{t=1}^{T(q)} a_{q,t}^R f_{q,t}e^{iq\varphi}. \quad (32)$$

Using the Fourier slice theorem and a few algebraic steps, it can be shown that the 2-D and the 3-D coefficients are related via

$$a_{q,t}^R = \sum_{\ell=|q|}^{L}\sum_{s=1}^{S(\ell)}\sum_{m=-\ell}^{\ell} A_{\ell,m,s}U_{m,q}^{\ell}(R)\gamma_{\ell,s}^{q,t}, \quad (33)$$

where $\gamma_{\ell,s}^{q,t}$ are constants depending on the radial functions:

$$\gamma_{\ell,s}^{q,t} = \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} Y_{\ell}^q(\frac{\pi}{2},\varphi)e^{-iq\varphi}F_{\ell,s}(\kappa)f_{q,t}(\kappa)\kappa d\kappa d\varphi. \quad (34)$$

In practice, the coefficients $\gamma_{\ell,s}^{q,t}$ are calculated via numerical integration over a closed segments [35], [48].

5) First Moment: By taking the expectation over $R \sim \rho$ and applying the distribution's expansion (30), we get

$$\mathbb{E}\left[a_{q,t}^R\right] = \sum_{\ell=|q|}^{\min(L,P)}\sum_{s=1}^{S(\ell)}\sum_{m=-\ell}^{\ell} A_{\ell,m,s}B_{\ell,-m,-q}\gamma_{\ell,s}^{q,t}\frac{(-1)^{m+q}}{2\ell+1}. \quad (35)$$

We note that the first moment is linear in both the volume's coefficients $\{A_{\ell,m.s}\}_{\ell,m.s}$ and the distribution $\{B_{p,u,v}\}_{p,u,v}$, as in the MRA model.

6) Second Moment: We have:

$$\mathbb{E}\left[a_{q_1,t_1}^R a_{q_2,t_2}^R\right] = \sum_{\substack{\ell_1,m_1,s_1, \\ \ell_2,m_2,s_2}} A_{\ell_1,m_1,s_1}A_{\ell_2,m_2,s_2}\gamma_{\ell_1,s_1}^{q_1,t_1}\gamma_{\ell_2,s_2}^{q_2,t_2}\times$$

$$\sum_p B_{p,-m_1-m_2,-q_1-q_2}C_p^{q_1,q_2}(\ell_1,\ell_2,m_1,m_2)\frac{(-1)^{m_1+m_2}}{2p+1}. \quad (36)$$

The first summation have the same range as in (35) for each set (ℓ_i,s_i,m_i). The second summation's range is

$$\max(|\ell_1-\ell_2|,|m_1+m+2|,|q_1+q+2|) \leq p \leq \min(\ell_1+\ell_2,P).$$

In addition:

$$C_p^{q_1,q_2}(\ell_1,\ell_2,m_1,m_2) =$$
$$C(\ell_1,m_1;\ell_2,m_2|p,m_1+m_2)C(\ell_1,q_1;\ell_2,q_2|p,q_1+q_2), \quad (37)$$

is the product of two Clebsch-Gordon coefficients. The second moment is a quadratic function of the volume's coefficients $\{A_{\ell,m.s}\}_{\ell,m.s}$ and is linear in the distribution's coefficients $\{B_{p,u,v}\}_{p,u,v}$, similarly to the MRA model.