GENERALIZED AUTOCORRELATION ANALYSIS FOR MULTI-TARGET DETECTION

Ye'Ela Shalit[†]**, Ran Weber*[†]**, Asaf Abas*[‡]**, Shay Kreymer*[†]**, and Tamir Bendory*[†]

[†]School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel
[‡]Department of Applied Mathematics, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

We study the multi-target detection problem of recovering a target signal from a noisy measurement that contains multiple copies of the signal at unknown locations. Motivated by the structure reconstruction problem in single-particle cryo-electron microscopy, we focus on the high noise regime, where the noise hampers accurate detection of the signal occurrences. We develop a generalized autocorrelation analysis framework to estimate the signal directly from the measurement. We demonstrate signal recovery from highly noisy measurements, and and show that the suggested framework outperform the classical autocorrelation analysis in a wide range of parameters. The code to reproduce all experiments is publicly available at `https://github.com/krshay/MTD-GMM`.

***Index Terms*—** Autocorrelation analysis, generalized method of moments, multi-target detection, cryo-electron microscopy.

1. INTRODUCTION

We study the multi-target detection (MTD) problem of estimating a target signal $x \in \mathbb{R}^L$ from a noisy measurement that contains multiple copies of the signal, each randomly translated [1], [2], [3], [4], [5], [6]. Specifically, let $y \in \mathbb{R}^N$ be a measurement of the form

$$y[\ell] = \sum_{i=1}^{p} x[\ell - \ell_i] + \varepsilon[\ell], \tag{1}$$

where $\{\ell_i\}_{i=1}^p \in \{L+1, \dots, N-L\}$ are arbitrary translations, and $\varepsilon[\ell] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. The translations and the number of occurrences of x in y, p, are unknown, and we assume $L \ll N$. Figure 1 presents examples of a measurement y at different signal-to-noise ratios (SNRs). We define SNR $:= \frac{\|x\|_2^2}{L\sigma^2}$.

The MTD model serves as a mathematical abstraction of the cryo-electron microscopy (cryo-EM) technology for macromolecular structure determination [7], [8], [9]. In a cryo-EM experiment [10], biological macromolecules suspended in a liquid solution are rapidly frozen into a thin ice layer. An electron beam then passes through the sample, and a two-dimensional tomographic projection is recorded. Importantly, the 2-D location and 3-D orientation of particles within the ice are random and unknown. This measurement is further affected by high noise levels and the optical configuration of the microscope.

In the current analysis workflow of cryo-EM data [11], [12], [13], the 2-D projections are first detected and extracted from the micrograph, and later rotationally and translationally aligned to reconstruct the 3-D molecular structure. This approach fails for small molecules, which induce low contrast, and thus low SNR. This makes them difficult to detect and align [6], [7], [14], rendering current cryo-EM algorithmic pipeline ineffective. For example, in the limit SNR $\to 0$, reliable detection of signals' locations within the measurement is impossible [6, Proposition 3.1].

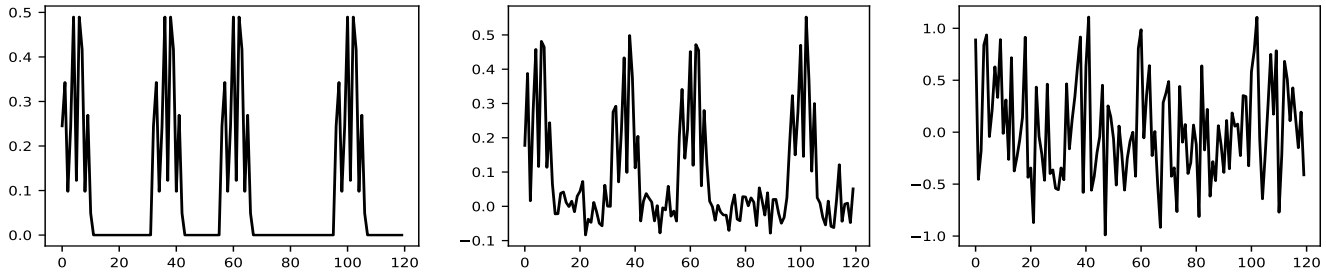
The MTD model was devised in [6] in order to study the recovery of small molecules directly from the micrograph, below the current detection limit of cryo-EM [7], [15]. An autocorrelation analysis technique (see Section 2.1) was implemented to recover low-resolution 3-D structures from noiseless simulated data under a simplified model. Autocorrelation analysis consists of finding a signal that best explains the empirical autocorrelations of the measurement, for example by minimizing a least-squares (LS) objective (see Section 2.3). For any noise level, those autocorrelations can be estimated to any desired accuracy for sufficiently large N. Computing the autocorrelations is straightforward and requires only one pass over the data, which is advantageous for massively large datasets, such as cryo-EM datasets [11].

Autocorrelation analysis is a special case of the method of moment, which is a classical statistical inference technique, tracing back to 1894 [16]. This work studies the application of the generalized method of moments [17], or, more precisely, the *generalized autocorrelation analysis*, to the MTD problem. In the classical autocorrelation analysis, we wish to find a signal whose autocorrelations match the observed autocorrelations. This is usually done by minimizing a LS objective. The generalized autocorrelation analysis suggests replacing the LS objective with a weighted LS, and provides a recipe of how to compute the optimal weights; see Section 2.4. The generalized method of moments has been proven to be highly effective in a variety of computational task, such as **Asaf, I guess you know what Tamir means here** [] We also mention that the framework can be formulated with other objective functions, rather than LS **and here.** [REF].

The main contribution of this paper is extending the autocorrelation analysis framework introduced in [1] by developing a generalized autocorrelation analysis framework for the one-dimensional MTD problem. We devise an algorithm for the recovery of the target signal from a measurement, and demonstrate a successful reconstruction in noisy regimes (see Section 3). Moreover, we show that the generalized autocorrelation analysis estimator outperforms the classical estimator in different SNRs and measurement's lengths (see Section 3). It is thus a first step towards applying the generalized method of moments to recovering small molecules from cryo-EM micrographs [6].

* These four authors have contributed equally to this work.

S.K. is supported by the Yitzhak and Chaya Weinstein Research Institute for Signal Processing. T.B. is supported in part by NSF-BSF grant no. 2019752.



(a) No noise. (b) SNR $= 50$. (c) SNR $= 0.1$.

Fig. 1: Three measurements at different noise levels: (a) no noise; (b) SNR $= 50$; (c) SNR $= 0.1$. Each measurement contains four copies of the target signal. In this work, our goal is to estimate the target signal directly from y. We focus on the low SNR regime (e.g., panel (c)) in which the signal occurrences are swamped by the noise, and the locations of the signal occurrences cannot be reliably detected.

2. COMPUTATIONAL FRAMEWORK

2.1. Autocorrelation analysis

The autocorrelation of order q of a signal $z \in \mathbb{R}^N$ is defined as

$$A_z^q[\ell_1, \ldots, \ell_{q-1}] := \mathbb{E}_z\Big[\frac{1}{N}\sum_{i\in\mathbb{Z}} z[i]z[i+\ell_1]\cdots z[i+\ell_{q-1}]\Big], \quad (2)$$

where $\ell_1, \ldots, \ell_{q-1}$ are integer shifts. Indexing out of bounds is zero-padded, that is, $z[i] = 0$ out of the range $\{0, \ldots, N-1\}$. In this work, we use the first three autocorrelations. As N grows indefinitely, the empirical autocorrelations of z almost surely (a.s.) converge to the population autocorrelations of z:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i\in\mathbb{Z}^2} z[i]z[i+\ell_1]\cdots z[i+\ell_{q-1}] \overset{\text{a.s.}}{=} A_z^q[\ell_1, \ldots, \ell_{q-1}], \quad (3)$$

by the law of large numbers.

Our goal is to relate the autocorrelations of the measurement y with the target signal x. In particular, the first-order autocorrelation is defined as

$$A_y^1 := \frac{1}{N}\sum_{i\in\mathbb{Z}} y[i], \quad (4)$$

which is just the mean of the measurement. The second-order autocorrelation of y, $A_y^2 : \mathbb{Z} \to \mathbb{R}$, is defined by

$$A_y^2[\ell_1] := \frac{1}{N}\sum_{i\in\mathbb{Z}} y[i]y[i+\ell_1], \quad (5)$$

and the third-order autocorrelation $A_y^3 : \mathbb{Z}\times\mathbb{Z} \to \mathbb{R}$ by

$$A_y^3[\ell_1, \ell_2] := \frac{1}{N}\sum_{i\in\mathbb{Z}} y[i]y[i+\ell_1]y[i+\ell_2]. \quad (6)$$

2.2. Autocorrelations under the well-separated model

We discuss the well-separated case of the MTD problem, which was studied in [1]. In this case, we assume that each signal in the measurement y is separated by at least a full signal length from its neighbors. Specifically, we assume that

$$|\ell_{i_1} - \ell_{i_2}| \geq 2L - 1, \quad \text{for all } i_1 \neq i_2. \quad (7)$$

Importantly, under the well-separated case, for shifts in the range $\mathcal{L} = \{0, \ldots, L-1\}$, any given occurrence of x in y is only ever correlated with itself, and never with another occurrence.

In [1], it was shown that under the well-separated condition (7), for any fixed level of noise σ^2, density γ and signal length L, in the limit $N \to \infty$, the autocorrelations of the measurements are equal, possibly up to some bias terms, to the autocorrelations of the signal times a density constant. Specifically, we have that

$$A_y^1 \overset{\text{a.s.}}{=} \gamma A_x^1, \quad (8)$$

$$A_y^2[\ell_1] \overset{\text{a.s.}}{=} \gamma A_x^2[\ell_1] + \sigma^2\delta[\ell_1], \quad (9)$$

$$A_y^3[\ell_1, \ell_2] \overset{\text{a.s.}}{=} \gamma A_x^3[\ell_1, \ell_2] + \gamma A_x^1\sigma^2(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2]), \quad (10)$$

for $\ell_1, \ell_2 \in \mathcal{L}$, where

$$\delta[\ell] = \begin{cases} 1 \text{ if } \ell = \vec{0}, \\ 0 \text{ otherwise}, \end{cases}$$

is the Kronecker delta function. Here, γ is the density of the target images in the measurement and is defined by

$$\gamma = p\frac{L}{N}. \quad (11)$$

Notably, the relations between the autocorrelations of y and x do not directly depend on the location of individual signal occurrences in the measurement, but only through the density parameter γ. Therefore, detecting the signal occurrences is not a prerequisite for signal recovery, and thus signal recovery is possible even in very low SNR regimes.

2.3. Signal recovery from autocorrelations

Following [1], [2], [3], [4], [5], and based on the notations presented above, we formulate a non-convex least squares problem for estimating the signal x from the autocorrelations of the measurement y:

$$\begin{aligned} \min_{x,\gamma>0} & w_1(A_y^1 - \gamma A_x^1)^2 \\ & + w_2\sum_{\ell_1=0}^{L-1} \|A_y^2[\ell_1] - \gamma A_x^2[\ell_1] - \sigma^2\delta[\ell_1]\|_2^2 \\ & + w_3\sum_{\ell_1=0}^{L-1}\sum_{\ell_2=0}^{L-1} \|A_y^3[\ell_1, \ell_2] - \gamma A_x^3[\ell_1, \ell_2] \\ & - \gamma A_x^1\sigma^2(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2])\|_2^2, \end{aligned} \quad (12)$$

where the weights were chosen such that each term is equally weighted, as suggested by [1]. This LS estimator serves as a benchmark for the generalized autocorrelation analysis estimator.

2.4. The generalizes autocorrelation analysis

2.4.1. The generalized autocorrelation analysis framework

In its most simplified form, the generalized autocorrelation analysis generalizes (12) by replacing the LS objective function with specific optimal weights. This choice guarantees favorable asymptotic statistical properties, such as the minimal asymptotic variance of the estimation error.

Let us define the *moment function*, $f(\theta, y)\colon \Theta \times \mathbb{R}^r \to \mathbb{R}^q$. The moment function is chosen such that its expectation value is zero only at a single point $\theta = \theta_0$, where θ_0 is the ground truth parameter set. Namely,

$$\mathbb{E}\left[f(\theta, y)\right] = 0 \quad \text{if and only if} \quad \theta = \theta_0. \tag{13}$$

The moment function must satisfy this uniqueness condition (13) and a few additional regularity conditions (which can be found in [17], [18] [19]). This flexibility enables the framework to be applied to a wide range of problems, such as subspace estimation [20].

In order to define the moment function for the MTD problem, we first define the i-th observation from the measurement y as follows:

$$y_i := [y[i], \ldots, y[i+L]]. \tag{14}$$

This choice of the moment function $f(\theta, y)$ must fulfill (13) using the defined samples y_i. The natural choice of $f(\cdot)$ is

$$f(\theta, y_i) := \begin{bmatrix} \gamma A_x^1 - A_{y_i}^1 \\ \left\{\gamma A_x^2[\ell_1] + \sigma^2\delta[\ell_1] - A_{y_i}^2[\ell_1]\right\}_{\ell_1=0}^{L-1} \\ \left\{\gamma A_x^3[\ell_1,\ell_2] + \gamma B(x,\ell_1,\ell_2) - A_{y_i}^3[\ell_1,\ell_2]\right\}_{\ell_1,\ell_2=0}^{L-1} \end{bmatrix}, \tag{15}$$

where $B(x,\ell_1,\ell_2) := A_x^1\sigma^2(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2])$, and $\theta := [x, \gamma]$. The estimated sample moment function is the average of f over N observations:

$$g_N(\theta) = \frac{1}{N}\sum_{i=0}^{N-1} f(\theta, y_i). \tag{16}$$

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

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta}\ g_N(\theta)^T W_N g_N(\theta). \tag{17}$$

Here, W_N is a fixed positive semi-definite (PSD) matrix. Note that the LS estimator (12) is a special case of (17), where W_N is the identity matrix.

2.4.2. 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

$$S := \lim_{N\to\infty} \mathrm{Cov}\left[\sqrt{N} g_N(\theta_0)\right], \tag{18}$$

be the covariance matrix of the estimated sample moment function (16) 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 matrix W. Finally, the expectation of the Jacobian of the moment function at the ground truth θ_0 is denoted by $G_0 = \mathbb{E}\left[\partial f(\theta_0, y)/\partial\theta^T\right]$.

The large sample properties of the GMM estimator were derived in [17], and are presented in the following theorem. The regularity conditions for this theorem can be found in [17], [18], [19].

Theorem 2.1. *Under the regularity conditions, 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 2.1 provides a matrix W that guarantees a minimal asymptotic variance of the estimator's error. The covariance matrix S of (18), which plays a central role in Theorem 2.1, is required to be a positive definite matrix. Therefore, the moment function must be chosen so that S is full-rank. As in [19], we remove the repeating entries of f (that appear due to the inherent symmetries of the autocorrelations). Moreover, in practice, the ground truth θ_0 is unknown a priori, so we cannot use the optimal weighting matrix. However, for our choice of the moment function (15), the matrix S is independent of the parameters of interest, and thus can be computed from the data:

$$\mathrm{Cov}[g(\theta)] = \mathrm{Cov}\left[\left\{[A_{y_i}^1; \mathcal{A}_{y_i}^2; \mathcal{A}_{y_i}^3]\right\}_{i=0}^{N-1}\right]. \tag{19}$$

3. NUMERICAL EXPERIMENTS

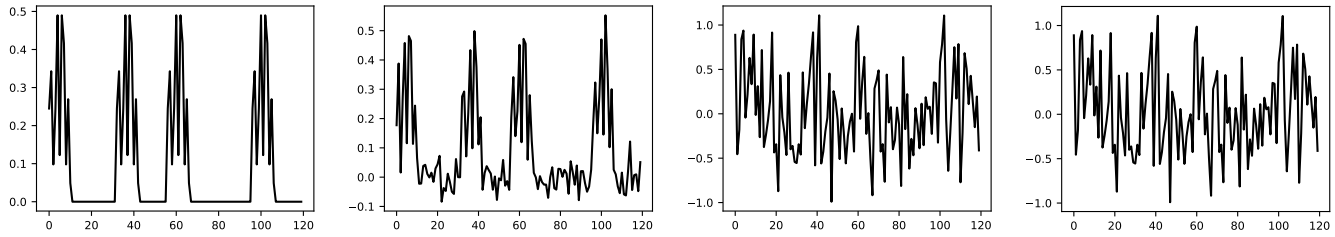
This section compares the numerical performance of the generalized autocorrelation analysis and the classical autocorrelation analysis. The optimization problem (12) was minimized using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, while ignoring the positivity constraint on γ (namely, treating it as an unconstrained problem). We measure the estimation error by

$$\mathrm{error}(x) = \frac{\|x - x^*\|_2}{\|x^*\|_2},$$

where x^* is the true signal, and x is the estimated signal. In experiments 3.2 and 3.3, the measurements were generated according to (1) with density $\gamma = 0.2$, and the target signals x are of length $L = 21$. Each entry in the signals was generated by $x[i] \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, 1)$, and the target signals were normalized such that $\|x\|_2 = 1$. We try to estimate the signal from 5 random initial guesses for x and from $\gamma_{\text{init}} = 0.18$, and calculate the estimation error for the estimate whose final objective function is minimal. Figures 3 and 4 present the median error over 50 trial. The code to reproduce all experiments is publicly available at `https://github.com/krshay/MTD-GMM`.

3.1. Recovery from a noisy measurement

In Figure 2 we present a successful recovery of a target signal of length $L = 11$ from a noisy measurement with SNR $= 1$ and $\gamma = 0.2$, using autocorrelation analysis and the generalized autocorrelation analysis. The noise level is visualized in Figure 1. As expected, the recovery error degrades as the measurement's length increases.



(a) $N = 1.5 \cdot 10^5$. (b) $N = 10^7$. (c) $N = 1.5 \cdot 10^8$. (d) Recovery errors.

Fig. 2: The recovery of the target signal x, as a longer and longer subset of y is observed, using autocorrelation analysis and the generalized autocorrelation analysis. Panel (d) presents the recovery errors using both methods, as a function of the measurement's length.

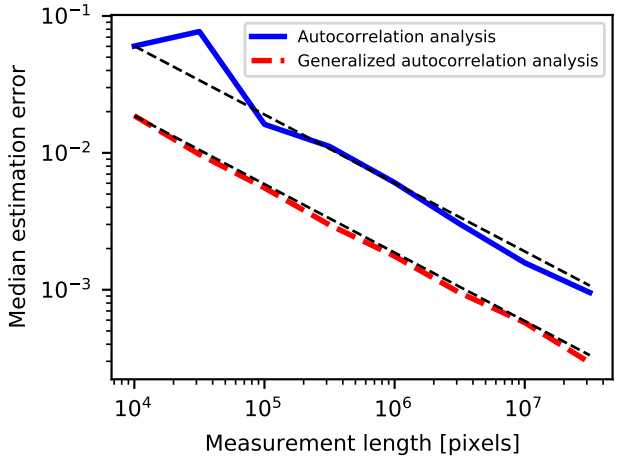


Fig. 3: The median estimation error of recovering the signal x, as a function of the measurement length N, by: (a) the autocorrelation analysis estimator; (b) the generalized autocorrelation analysis estimator. Evidently, the generalized autocorrelation analysis significantly outperforms the classical autocorrelation analysis for all N.

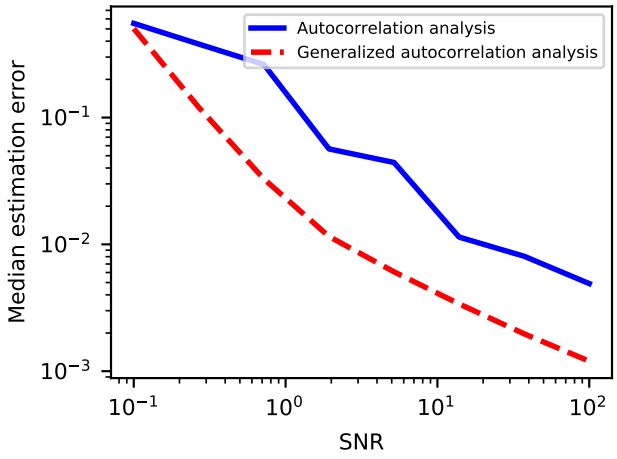


Fig. 4: The median estimation error of recovering the signal x, as a function of SNR, by: (a) the autocorrelation analysis estimator; (b) the generalized autocorrelation analysis estimator.

3.2. Recovery error as a function of the measurement length

Figure 3 presents recovery error as a function of the measurement length N. We set SNR $= 50$. As expected from the law of large numbers, the recovery error of both estimators decays as $N^{-1/2}$, where N is the length of the measurement. Evidently, the generalized autocorrelation analysis significantly outperforms the classical autocorrelation analysis for all N.

3.3. Recovery error as a function of SNR

Figure 4 presents recovery error as a function of SNR, for $N = 10^6$. For all levels of SNR, the recovery error using the GMM is smaller than the error using the regular MoM estimator. In addition, the slope of the error curve increases dramatically at low SNR, which is a known phenomenon in the cryo-EM literature, see for example [21], [22], [23].

4. CONCLUSION

This paper is motivated by the effort of reconstructing small 3-D molecular structures using cryo-EM, below the current detection limit. The main contribution of this study is incorporating the generalized method of moments into the computational framework proposed in [6] for the MTD problem. The optimality of the generalized framework was presented in Theorem 2.1, and corroborated by numerical experiments. Previous numerical experiments [5] suggest that naive detection, extraction and averaging fails for low SNR regimes, such that study of methods that do not require accurate detection of signal occurrences is of significance to the problem on hand.

Future work includes extending the generalized autocorrelation analysis estimator to the 2-D and 3-D cases of the MTD problem, and extending the framework to the case of an arbitrary spacing distribution between signal and image occurrences [2], [5]. Designing other inference techniques, such as maximum likelihood estimators [2], for the two-dimensional MTD problem is also a possible future research direction.

5. REFERENCES

[1] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer, "Multi-target detection with application to cryo-electron microscopy," *Inverse Problems*, vol. 35, no. 10, pp. 104003, 2019.

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

[3] Nicholas F Marshall, Ti-Yen Lan, Tamir Bendory, and Amit Singer, "Image recovery from rotational and translational invariants," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5780–5784.

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

[5] Shay Kreymer and Tamir Bendory, "Two-dimensional multi-target detection: an autocorrelation analysis approach," *arXiv preprint arXiv:2105.06765*, 2021.

[6] 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.

[7] Richard Henderson, "The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules," *Quarterly Reviews of Biophysics*, vol. 28, no. 2, pp. 171–193, 1995.

[8] Eva Nogales, "The development of cryo-EM into a mainstream structural biology technique," *Nature methods*, vol. 13, no. 1, pp. 24–27, 2016.

[9] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in Biochemical Sciences*, vol. 40, no. 1, pp. 49–57, 2015.

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

[11] Tamir Bendory, Alberto Bartesaghi, and Amit Singer, "Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities," *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 58–76, 2020.

[12] Sjors HW Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519–530, 2012.

[13] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker, "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination," *Nature methods*, vol. 14, no. 3, pp. 290–296, 2017.

[14] Cecilia Aguerrebere, Mauricio Delbracio, Alberto Bartesaghi, and Guillermo Sapiro, "Fundamental limits in multi-image alignment," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5707–5722, 2016.

[15] Edoardo D'Imprima and Werner Kühlbrandt, "Current limitations to high-resolution structure determination by single-particle cryoEM," *Quarterly Reviews of Biophysics*, vol. 54, 2021.

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

[17] Lars Peter Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica*, vol. 50, no. 4, pp. 1029, 1982.

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

[19] Asaf Abas, Tamir Bendory, and Nir Sharon, "The generalized method of moments for multi-reference alignment," *arXiv preprint arXiv:2103.02215*, 2021.

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

[21] Fred J Sigworth, "A maximum-likelihood approach to single-particle image refinement," *Journal of structural biology*, vol. 122, no. 3, pp. 328–339, 1998.

[22] 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*, vol. 65, no. 6, pp. 3565–3584, 2018.

[23] 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*, vol. 1, no. 3, pp. 497–517, 2019.