MULTI-SEGMENT RECONSTRUCTION USING INVARIANT FEATURES

Mona Zehni, Minh N. Do, Zhizhen Zhao

Department of ECE and CSL, University of Illinois at Urbana-Champaign

ABSTRACT

Multi-segment reconstruction (MSR) problem consists of recovering a signal from noisy segments with unknown positions of the observation windows. One example arises in DNA sequence assembly, which is typically solved by matching short reads to form longer sequences. Instead of trying to locate the segment within the sequence through pair-wise matching, we propose a new approach that uses shift-invariant features to estimate both the underlying signal and the distribution of the positions of the segments. Using the invariant features, we formulate the problem as a constrained nonlinear least-squares. The non-convexity of the problem leads to its sensitivity to the initialization. However, with clean data, we show empirically that for longer segment lengths, random initialization achieves exact recovery. Furthermore, we compare the performance of our approach to the results of expectation maximization and demonstrate that the new approach is robust to noise and computationally more efficient.

Index Terms— multi-segment reconstruction, invariant features, non-convex optimization, DNA sequence assembly, cryo-EM

1. INTRODUCTION

We consider the following observation model,

$$y_k = \mathcal{M}_{s_k} x + \varepsilon_k, \quad k \in \{1, 2, ..., K\} \tag{1}$$

where $y_k \in \mathbb{R}^m$ and $x \in \mathbb{R}^d$, $m \leq d$, correspond to the k-th observation and the underlying signal respectively. $\mathcal{M}_s$ denotes a cyclic masking operator that captures m consecutive entries of the signal starting from location s. In other words, $\mathcal{M}_s : \mathbb{R}^d \to \mathbb{R}^m$ and $(\mathcal{M}_s x)[n] = x[n + s \bmod d]$. For the sake of brief notations, we define $x[n+s]_d := x[n+s \bmod d]$ from now on. We also assume $s \in \{0, 1, ..., d-1\}$ to be a random variable drawn from a general distribution with p as its probability mass function, i.e. $P\{s = s_k\} = p[s_k]$. Furthermore, the randomly located segment of the signal is contaminated by additive white Gaussian noise ε_k with zero mean and variance $\sigma^2 I_m$. Figure 1 further illustrates the observation model (1).

Our goal here is to recover x from noisy partial observations $\{y_k\}_{k=1}^K$. This problem is linked to *multi-reference alignment (MRA)* [1] in which estimation of the signal from noisy and random circularly-shifted versions of itself is targeted [2, 3]. While in MRA the whole signal takes part in each observation, in MSR shorter segments of the signal contribute to the observations. Similar problems to MSR appear in DNA sequencing [4, 5], common superstring problem [6, 7, 8], puzzle solving [9], image registration, super-resolution imaging [10] and cryo-EM [11, 12], to name a few.

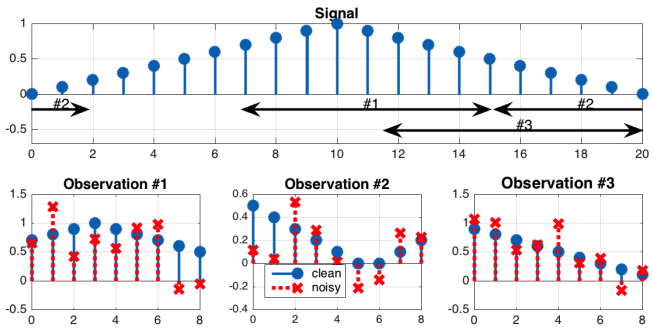

Fig. 1. (a) The original signal, (b) several noisy circular segments of the signal

MSR is originally motivated by DNA sequencing, short common super string (SCS) problem and cryo-EM. While in DNA sequencing, assembling the whole sequence from short-length reads is addressed, in SCS finding the shortest string containing a set of strings is the ultimate target. Also, our signal model is relevant to the cryo-EM 3D reconstruction in the sense that the Fourier transform of a 2D projection image is a partial observation of the 3D volume.

In this paper, we have assumed fixed segment length, however the same approach can be easily extended to random-length segments. In addition, we assume that the probability mass function of the positions of the segments is no longer uniform unlike [4], thus further generalizing the problem. Next, we estimate some features of the signal from the observations and then try to recover the signal from those features. The advantages of pursuing this approach compared to using the observations directly are in *1)* we estimate x and p directly, thus circumventing the estimation of $\mathcal{M}_{s_k}$ which is most of the times impossible due to high level of noise [13]-[14], *2)* creating features that are shift-invariant, i.e. if x and p are shifted the same amount, the constructed features will not change, *3)* instead of using probabilistic models such

as maximum-likelihood which are computationally expensive [13, 15, 16], our approach goes through the observations once to create invariant features similar to [2, 11, 17], *4)* the signal recovery is more robust to noise as the features can be estimated accurately when sufficient number of observations is available.

We formulate the problem of recovering x and p as a weighted non-linear least-squares problem. We seek to find a signal that matches the higher order statistics (up to third order correlation) derived from the observations. In addition, due to the structure of the constructed features, the formulated optimization problem can be viewed as a tensor decomposition problem [18]-[19]. Our simulations reveal that this problem has spurious local minima apart from the global minima, hence additional care should be given to the initialization scheme. Note that x and p are only determined up to a global cyclic shift. We can clearly see that as the length of the segment increases, the convergence of the recovery problem to the accurate solution becomes less sensitive to random initialization. Additionally, we observe that it is impossible to recover from partial observations of the signal that are shorter than a threshold, similar to [4]. We also apply our approach to some small scale form of gene sequencing problem. Relying on the results we hope we can further extend our problem to larger scales such that it finds real applications in DNA sequence assembly. MATLAB implementations of our paper are provided in https://github.com/MonaZI/MSR.

The organization of the paper is as follows. In Section 2 we describe our method. In Section 3 we present the results of our approach and finally conclude the paper in Section 4.

2. METHODS

We use the first, second, and third order correlation of the signal as the shift-invariant features. Let μ, C and T be the population expectations corresponding to these features obtained from clean observations as in (2).

$$\mu_{x,p}[n] = \sum_{s=0}^{d-1} x[n+s]_d \, p[s], \tag{2}$$

$$C_{x,p}[n_1,n_2] = \sum_{s=0}^{d-1} x[n_1+s]_d \, x[n_2+s]_d \, p[s],$$

$$T_{x,p}[n_1,n_2,n_3] = \sum_{s=0}^{d-1} x[n_1+s]_d \, x[n_2+s]_d \, x[n_3+s]_d \, p[s].$$

We use the observations y_k to construct empirical estimates of invariants in (2) as in (3). [2] verifies that the relative error in the estimation of the second and third-order correlation decays as $\frac{1}{\sqrt{K}}$ and the sample complexity for the estimation of T and C is $O(\sigma^6)$ and $O(\sigma^4)$ respectively.

$$\widehat{\mu}[n] = \frac{1}{K}\sum_{k=1}^{K} y_k[n] \to \mu_{x,p}[n], \tag{3}$$

$$\widehat{C}[n_1,n_2] = \frac{1}{K}\sum_{k=1}^{K} y_k[n_1]y_k[n_2] - \sigma^2\delta(n_1,n_2) \to C_{x,p}[n_1,n_2]$$

$$\widehat{T}[n_1,n_2,n_3] = \frac{1}{K}\sum_{k=1}^{K} y_k[n_1]y_k[n_2]y_k[n_3] - \sigma^2\,(\widehat{\mu}[n_1]\delta(n_2,n_3)$$

$$+\widehat{\mu}[n_2]\delta(n_1,n_3) + \widehat{\mu}[n_3]\delta(n_1,n_2)) \to T_{x,p}[n_1,n_2,n_3]$$

Thus, MSR formulation for non-uniform p is described in (4) where $\|.\|_F$ marks the Frobenius norm,

$$\min_{x,p} \lambda_T\|\widehat{T} - T_{x,p}\|_F^2 + \lambda_C\|\widehat{C} - C_{x,p}\|_F^2 + \lambda_\mu\|\widehat{\mu} - \mu_{x,p}\|_2^2$$

$$s.t. \quad \forall i \in \{0,\dots,d-1\}, p[i] \geq 0, \ \sum_{i=0}^{d-1} p[i] = 1. \tag{4}$$

In case of uniform distribution for s, $p[s=i] = \frac{1}{d}, \forall\, i \in \{0,1,...,d-1\}$, the dimension of the invariant features will further reduce due to existing symmetries. As a result, similar derivations to (2) simplify as,

$$\widetilde{\mu}_x = \frac{1}{d}\sum_{m=0}^{d-1} x\,[m]\,, \widetilde{C}_x\,[n] = \frac{1}{d}\sum_{m=0}^{d-1} x[m+n]_d x[m],$$

$$\widetilde{T}_x\,[n_1,n_2] = \frac{1}{d}\sum_{m=0}^{d-1} x[n_1+m]_d x[n_2+m]_d x[m]. \tag{5}$$

The MSR formulation for uniform p is,

$$\min_{x} \lambda_T\|\widehat{T} - \widetilde{T}_x\|_F^2 + \lambda_C\|\widehat{C} - \widetilde{C}_x\|_F^2 + \lambda_\mu\|\widehat{\mu} - \widetilde{\mu}_x\|_2^2 \tag{6}$$

where we reuse $\widehat{T}$, $\widehat{C}$ and $\widehat{\mu}$ notations to also refer to the empirical estimates of $\widetilde{T}_x$, $\widetilde{C}_x$ and $\widetilde{\mu}_x$ respectively. Although (6) is a special case of (4) there are a few differences that makes the former interesting to study separately. First, the only unknown we seek to recover in (6) is the signal x unlike (4) in which both x and p are undetermined. Also, (6) is an unconstrained optimization problem, while (4) requires p to be a valid discrete probability mass function. Besides, the complexity of (6) is further reduced due to lower dimensions of the invariant features.

Note that the objective functions in both problems correspond to the weighted squared Frobenius distance between the ground truth features of x and their estimated values. λ_T, λ_C and λ_μ are the weights we give to the importance of matching the third, second and first order correlation terms with their estimated values. For example, as λ_T increases, x and p are set in a way that further match $\widehat{T}$. In Section 3, we also examine the case with $\lambda_T = 0$ to see the possibility of

accurate recovery of x and p by merely using statistics up to second order.

Note the objective function in (4) which is a 6-th order polynomial in x and 2-nd order polynomial in p. Thus, although the constraints form a convex set, the overall optimization problem is non-convex. There are couple of challenges with non-convex optimization, *1)* existence of local minima and *2)* existence of saddle points which might slow down the first-order optimization approaches. To avoid the second pitfall, we exploit second order methods such as trust-region and sequential quadratic programming (SQP) [20]-[21], implemented in MATLAB optimization toolbox.

In addition to the local non-convex optimization approach, we use global optimization with polynomials [22] to reconstruct the signal from the invariant moments. More specifically, the objective function is a sum of squares (SOS) polynomial. The Lasserre hierarchy of relaxations is able to solve the MSR problem for small d with m above a d-dependent threshold[1]. However, it becomes computationally expensive and requires too much memory for $d > 9$.

2.1. Analysis

Here we briefly analyze our problem for the clean case with $\sigma = 0$. It is worth mentioning that as the simultaneous shifts of x and p result in the same features, there are at least d global minima. When the segment length is small, the number of algebraically independent equations provided by the invariant features in (2) is not enough to uniquely determine x and p. Therefore, random initialization with local non-convex algorithms are able to achieve the global minima, but fail to recover the true signal and the corresponding segment location pmf.

Let us denote $\tilde{m}(d)$ as the minimum m for which the number of algebraically independent equations provided by the invariant features reaches the number of unknowns. $\tilde{m}(d)$ varies across different problem settings as,

$$\tilde{m}(d) = \min_{m \in \mathbb{N}} m$$

$$s.t. \quad \begin{cases} \frac{m^3}{6} + m^2 + \frac{11}{6}m + 1 \geq 2d & \text{non-unif. } p, \lambda_T \neq 0 \\ \frac{m^2}{2} + \frac{3}{2}m + 1 \geq 2d & \text{non-unif. } p, \lambda_T = 0 \\ \frac{m^2}{2} + \frac{3}{2}m + 1 \geq d & \text{unif. } p, \lambda_T \neq 0 \end{cases} \tag{7}$$

We provide numerical results to verify our analysis of $\tilde{m}(d)$. In addition, we can extend our approach to contain moments up to $s > 3$ and count the number of algebraically independent equations in order to determine the minimum required segment length.

In bispectrum inversion for 1D MRA [2], it is observed that with random initialization, local non-convex algorithm is able to exactly recover the signal in the noiseless case. However, for MSR, $m > \tilde{m}(d)$ is not sufficient to guarantee the exact recovery from random initialization. In some cases, gradient methods can get stuck in the local minima and therefore require good initialization. Similar phenomenon is observed in [23] for reconstructing heterogeneous signals from invariant moments.

[1]This will be further discussed in the sequel.

3. NUMERICAL RESULTS

In our simulations we generate x and p randomly. Also, we adopt 10^5 noisy observations to estimate the shift-invariant features as in (3). Also, in (4) and (6) we assume $\lambda_\mu = \lambda_C = \lambda_T = 1$ unless otherwise stated. To assess our methodology we use several performance metrics, *1)* mean-squared error defined as MSE $= \|x - \hat{x}\|^2$, *2)* the probability of accurate recovery, i.e. $p_{rec}(th) = \text{P}\{\text{MSE} \leq th\}$, *3)* the median of the final value of the objective function denoted by $\overline{f}$ which is an indicator of whether the globally optimal solution is obtained. For our evaluations in this section, we set $th = 10^{-3}$. To derive $p_{rec}(th)$ and $\overline{f}$, we solve the optimization problem using trust-region and SQP starting from a random initial point for 50 trials. Note that when we state accurate recovery is achieved, we mean an accurate estimation of x and p is recovered up to a global cyclic shift and the corresponding MSE is below th. In what follows, we discuss the two main results of our experiments.

• *The impact of the segment length on the possibility of getting to the global minima*: We investigate the changes of $p_{rec}(th)$ and $\overline{f}$ with respect to m and d for four different cases when $\sigma = 0$ as illustrated in Fig. 2, *a)* uniform p and $\lambda_T \neq 0$, *b)* non-uniform p and $\lambda_T = 0$, *c)* non-uniform p and $\lambda_T \neq 0$, *d)* non-uniform p, $\lambda_T \neq 0$ and x discretized in value, i.e. $x[n] \in \{0, 1, 2, 3\}, \forall n \in \{0, 1, ..., d-1\}$.

An immediate observation from all four subplots in Fig. 2 suggests that the larger the m, the higher $p_{rec}(th)$. Comparing Fig. 2(a) with Fig. 2(b) verifies that for uniform p, the minimum length of segments required for accurate recovery is smaller compared to non-uniform p and $\lambda_T = 0$ case. Also, for non-uniform p when $\lambda_T \neq 0$, $\tilde{m}(d)$ is smaller compared to the case of $\lambda_T = 0$, as also predicted by (7). This clearly proposes that using the 3-rd order correlation provides more information about the signal and thus accurate signal recovery can be obtained for smaller m. Additionally, Fig. 2(d),(h) shows how our proposed method can be extended to problems in which the signal is discretized in value (similar to a DNA sequence assembly problem) and again how accurate recovery is achievable when the length of the reads surpass a certain threshold.

Regarding the landscape of the problem what we observe is, *1)* for small values of m the global minimum is not unique (up to cyclic shifts) and reaching global minima does not necessarily guarantee accurate reconstruction and, *2)* the problem has local minima. The evidence for the first statement is that for some trials, although the value of the objective function at the optimal point is reported very small ($\sim 10^{-10}$), $\hat{x}$ and

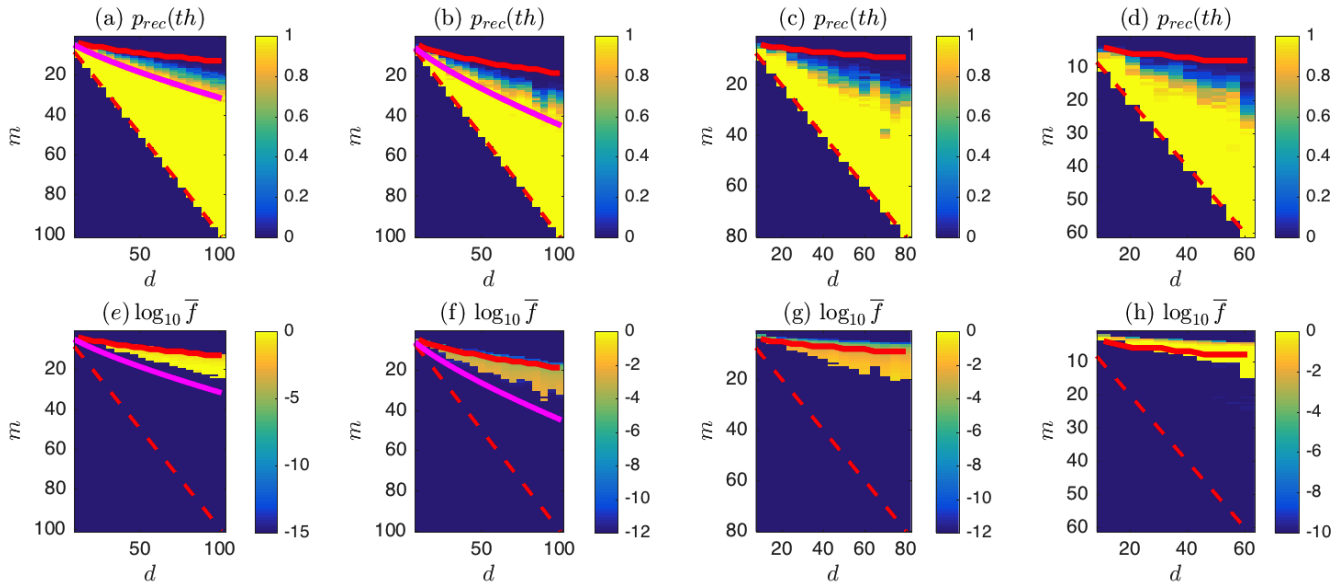

Fig. 2. $p_{rec}(th)$ and $\overline{f}$ for *a,e)* uniform p and $\lambda_T \neq 0$ in (6), *b,f)* non-uniform p and $\lambda_T = 0$ in (4), *c,g)* non-uniform p and $\lambda_T \neq 0$ in (4), *d,h)* x has discrete values, non-uniform p and $\lambda_T \neq 0$ in (4). The red solid lines mark $\tilde{m}(d)$, the red-dashed lines convey the upper bound on m, i.e. $m \leq d$ and the solid magenta line locate the minimum m for each d for which $p_{rec}(th)$ becomes one. For *(a,e)* and *(b,f)* the magenta line is fit to $d^{\frac{3}{4}}$ and $\sqrt{2}d^{\frac{3}{4}}$ respectively.

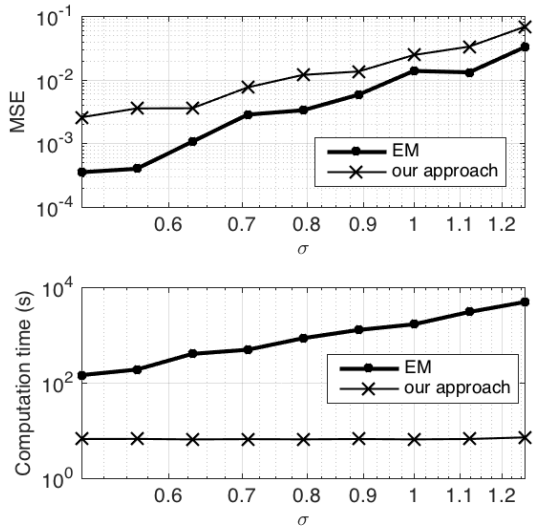

Fig. 3. The comparison between the results of our approach and EM in terms of MSE and computation time for different noise levels and fixed $d = 45$ and $m = 25$.

$\hat{p}$, do not match their true values, as displayed in Fig. 2. The $d - m$ region for which this happens is the blue-colored region on top of the yellow strip in Fig. 2(e)-(h) which almost maps to $m < \tilde{m}(d)$ region. Additionally, we noticed that in some trials, when reaching the local minimum is reported with relatively larger values of the objective function at the optimal point, the solution does not match the original x and p. This also marks the existence of local minima in addition to global minima. The corresponding $d - m$ region for this case is marked by the yellow shaded regions in Fig. 2(e)-(h).

• *Robustness of the recovery to noise and comparison with expectation-maximization (EM) method*: Figure 3 compares the performance of our approach with the results obtained from expectation maximization [24] for different noise levels. It can be inferred that in high noise regimes the performance of both our approach and EM degrades. On the other hand, our approach is computationally more efficient and scales linearly with the number of samples, so it can be used as a good initialization for EM.

4. CONCLUSION

In this paper, we proposed a new approach for recovering a signal from a large number of randomly observed noisy segments. The random locations of the observation windows are unknown. Instead of trying to recover the locations for each segment through matching, we used shift invariant features to estimate the underlying signal and the distribution of the windows. The invariant features approach has low computational complexity for large sample size compared to alternative methods, such as EM. The signal is reconstructed by solving a constrained nonlinear least-squares problem. Due to the non-convex nature of the problem, the solution depends on the initialization. It was shown that for clean data, as the length of the segment increases, random initialization can achieve accurate recovery. We also demonstrated that the new method is robust to noise and efficient in terms of computational time.

5. REFERENCES

[1] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu, "Multireference alignment using semidefinite programming," in *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM, 2014, pp. 459–470.

[2] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, "Bispectrum Inversion with Application to Multireference Alignment," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 1037–1050, 2017.

[3] E. Abbe, T. Bendory, W. Leeb, J. Pereira, N. Sharon, and A. Singer, "Multireference alignment is easier with an aperiodic translation distribution," *arXiv preprint arXiv:1710.02793*, Oct. 2017.

[4] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Transactions on Information Theory*, vol. 59, pp. 6273 – 6289, 2013.

[5] E. D. Green, "Strategies for the systematic sequencing of complex genomes," *Nature Reviews, GENETICS*, vol. 2, pp. 573–583, 2001.

[6] A. Frieze and W. Szpankowski, "Greedy algorithms for the shortest common superstring that are asymptotically optimal," *Algorithmica*, vol. 21, no. 1, pp. 21–36, 1998.

[7] Haim Kaplan and Nira Shafrir, "The greedy algorithm for shortest superstrings," *Information Processing Letters*, vol. 93, no. 1, pp. 13–17, 2005.

[8] Bin Ma, "Why greed works for shortest common superstring problem," *Theoretical Computer Science*, vol. 410, no. 51, pp. 5374–5381, 2009.

[9] G. Paikin and A. Tal, "Solving multiple square jigsaw puzzles with missing pieces," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 161–174.

[10] P. Vandewalle, L. Sbaiz, J. Vandewalle, and M. Vetterli, "Super-resolution from unregistered and totally aliased signals using subspace methods," *IEEE Transactions on Signal Processing*, vol. 55, pp. 3687 – 3703, 2007.

[11] Z. Kam and I. Gafni, "Three-dimensional reconstruction of the shape of human wart virus using spatial correlations," *Ultramicroscopy*, vol. 17, pp. 251–262, 1985.

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

[13] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak, "Rapid solution of the cryo-em reconstruction problem by frequency marching," *SIAM Journal on Imaging Sciences*, vol. 10, no. 3, pp. 1170–1195, 2017.

[14] V. L. Shneerson, A. Ourmazd, and D. K. Saldin, "Crystallography without crystals. I. the common-line method for assembling a three-dimensional diffraction volume from single-particle scattering," *Acta Crystallographica*, 2008.

[15] A. Punjani, M. A. Brubaker, and D. J. Fleet, "Building proteins in a day: Efficient 3D molecular structure estimation with electron cryomicroscopy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 706–718, 2017.

[16] S.H.W. Scheres, "Chapter Six - Processing of structurally heterogeneous cryo-EM data in RELION," *Methods in Enzymology*, vol. 579, pp. 125–157, 2016.

[17] K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doeleman, and W. T. Freeman, "Computational Imaging for VLBI Image Reconstruction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2016.

[18] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM REVIEW*, vol. 51, no. 3, pp. 455–500, 2009.

[19] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab 3.0," Mar. 2016, Available online.

[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, 2006.

[21] S. J Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, and A. J Smola, "A Generic Approach for Escaping Saddle points," *arXiv preprint arXiv:1709.01434*, Sept. 2017.

[22] Jean B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 796–817, 2001.

[23] N. Boumal, T. Bendory, R. R. Lederman, and A. Singer, "Heterogeneous multireference alignment: a single pass approach," *arXiv preprint arXiv:1710.02590*, Oct. 2017.

[24] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.