Fundamental Limits in Multireference Alignment

Nir Sharon

PACM, Princeton University

October 11, 2017
1. The Problem of MultiReference Alignment (MRA)

2. Moments Approach and Theoretical Guarantees

3. Numerical Examples and Conclusions
Table of contents

1. The Problem of MultiReference Alignment (MRA)

2. Moments Approach and Theoretical Guarantees

3. Numerical Examples and Conclusions
The “resolution revolution”  
(Werner Kühlbrandt, 2014)
Exciting Times:
“For developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution”
The process:
Motivational Section – Cryo-EM in a Nutshell

The process:

Main computational challenges:
1. High level of noise.
2. Unknown viewing directions.
The Problem of Multi-Reference Alignment (MRA)

**Formulation:** Estimating a signal $x \in \mathbb{R}^L$, up to shifting, from its noisy circularly–translated copies

$$y_j = R_{r_j} x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

**Applications:** radar, image registration, structural biology
The Sample Complexity of the Problem

**Data:**

![Data Diagram]
Can we recover the signal for any level of noise?
How many samples do we need for attaining a certain accuracy?
The Sample Complexity of the Problem

Data:

![Graphs of data with different scales and N values]

Estimations:

- $N = 1,000$
- $N = 100,000$
- $N = 10,000,000$
Recall the model:

\[ y_j = R_j x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I). \]

- The translations are the latent/hidden variables of the problem.
The Role of Translations in MRA

Recall the model:

\[ y_j = R_{r_j}x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I). \]

- The translations are the latent/hidden variables of the problem.
- Given the translations, we can estimate

\[
\tilde{x} = \frac{1}{N} \sum_{j=1}^{N} R_{r_j}^{-1} y_j.
\]

Therefore, estimating the translation reduces the problem significantly.
The Role of Translations in MRA

Recall the model:

$$y_j = R_{r_j} x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

- The translations are the latent/hidden variables of the problem.
- Given the translations, we can estimate

$$\hat{x} = \frac{1}{N} \sum_{j=1}^{N} R_{r_j}^{-1} y_j.$$

Therefore, estimating the translation reduces the problem significantly.

1. Should we find the translations to estimate the signal?
2. Can we recover the signal without explicit translations?
Low vs High Level of Noise

An example – alignment via cross-correlation

Observation 1
Observation 2
Cross-correlation

\[ \leq 0 \]
\[ \leq 0.1 \]
\[ \leq 0.7 \]
\[ \leq 2 \]
Low vs High Level of Noise

- Low level of noise $\rightarrow$ Explicitly estimating the translations.
Low vs High Level of Noise

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

An example – alignment via cross-correlation
Low vs High level of noise Regimes

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

- High level of noise $\rightarrow$ Keeping translations behind the scene:
Low vs High level of noise Regimes

- Low level of noise → Explicitly estimating the translations.

- High level of noise → Keeping translations behind the scene:
  - Invariant features – the info is encompassed in mean, power spectrum, and bispectrum.
  - EM – a classical statistical method with state-of-the-art performance for MRA.
Low vs High level of noise Regimes

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

- High level of noise $\rightarrow$ Keeping translations behind the scene:
  - Invariant features – the info is encompassed in mean, power spectrum, and bispectrum.
  - EM – a classical statistical method with state-of-the-art performance for MRA.

- The sample complexity tradeoff,

$$N \gtrsim \sigma^2 \quad \text{versus} \quad N \gtrsim \sigma^6$$
Table of contents

1. The Problem of MultiReference Alignment (MRA)

2. Moments Approach and Theoretical Guarantees

3. Numerical Examples and Conclusions
The first moment of the data is
\[
E[y] = C_x \rho = x \ast \rho,
\]
where \( C_x \) is a circulant matrix and \( \rho \) is the distribution of translations.

The second moment of the data is
\[
E[yy^T] = C_x D \rho C_x^T + \sigma^2 I,
\]
where \( D \rho \) is a diagonal matrix with \( \rho \) on its diagonal.

Proposition
Assume the DFT of \( \rho \) satisfies \( \hat{\rho}[k] \neq 0 \) for some \( k \), where \( k \) and \( L \) are coprime. If the DFT of \( x \) is non-vanishing, then it is uniquely determined (up to translation) from the first two moments of the data.
The first moment of the data is

$$E[y] = C_x \rho = x \ast \rho,$$

where $C_x$ is a circulant matrix and $\rho$ is the distribution of translations.
The first moment of the data is
\[ \mathbb{E} [y] = C_x \rho = x * \rho, \]
where \( C_x \) is a circulant matrix and \( \rho \) is the distribution of translations.

The second moment of the data is
\[ \mathbb{E} [yy^T] = C_x D_{\rho} C_x^T + \sigma^2 I, \]
where \( D_{\rho} \) is a diagonal matrix with \( \rho \) on its diagonal.
The first moment of the data is

$$\mathbb{E}[y] = C_x \rho = x \ast \rho,$$

where $C_x$ is a circulant matrix and $\rho$ is the distribution of translations.

The second moment of the data is

$$\mathbb{E}[yy^T] = C_x D_{\rho} C_x^T + \sigma^2 I,$$

where $D_{\rho}$ is a diagonal matrix with $\rho$ on its diagonal.

**Proposition**

Assume the DFT of $\rho$ satisfies $\hat{\rho}[k] \neq 0$ for some $k$, where $k$ and $L$ are coprime. If the DFT of $x$ is non-vanishing, then it is uniquely determined (up to translation) from the first two moments of the data.
The first moment of the data is

$$\mathbb{E} [y] = C_x \rho = x \ast \rho,$$

where $C_x$ is a circulant matrix and $\rho$ is the distribution of translations.

The second moment of the data is

$$\mathbb{E} [yy^T] = C_x D_\rho C_x^T + \sigma^2 I,$$

where $D_\rho$ is a diagonal matrix with $\rho$ on its diagonal.

Assume $|\hat{x}[k]| = 1$, we have

$$C_x D_\rho C_x^T = C_x D_\rho C_x^{-1},$$

and $x$ is recovered whenever $\rho$ has a distinct entry.
The success of the spectral algorithm is contingent on a distinct entry in $\rho$. Can we somehow guarantee it?

Proposition

Let $\rho$ be a non-periodic vector on the simplex and let $\theta$ be a random probability density function on the simplex. Then, all entries of $\rho^* \theta$ are distinct with probability $1$.

A conclusion: the spectral algorithm provides a constructive solution for any non-periodic $\rho$ and achieves a sample complexity of $N \gtrsim \sigma^4$ (equivalently of order $1/\text{SNR}^2$).
The success of the spectral algorithm is contingent on a distinct entry in $\rho$. Can we somehow guarantee it?

The answer: Yes! by randomly reshuffle the samples!
The success of the spectral algorithm is contingent on a distinct entry in $\rho$. Can we somehow guarantee it?

The answer: Yes! by randomly reshuffle the samples!

**Proposition**

Let $\rho$ be a non-periodic vector on the simplex and let $\theta$ be a random probability density function on the simplex. Then, all entries of $\rho \ast \theta$ are distinct with probability 1.
The success of the spectral algorithm is contingent on a distinct entry in $\rho$. Can we somehow guarantee it?

The answer: Yes! by randomly reshuffle the samples!

**Proposition**

Let $\rho$ be a non-periodic vector on the simplex and let $\theta$ be a random probability density function on the simplex. Then, all entries of $\rho \ast \theta$ are distinct with probability 1.

**A conclusion:** the spectral algorithm provides a constructive solution for any nonperiodic $\rho$ and achieves a sample complexity of $N \gtrsim \sigma^4$ (equivalently of order $1/\text{SNR}^2$).
The Spectral Algorithm in Full

1. Estimate moments and power spectrum
   \[ \mu \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j \]
   \[ \sigma^2 \]

2. Normalize the signal
   \[ \bar{p} \leftarrow (\bar{p})_{-1/2} \]
   \[ \bar{q} \leftarrow \bar{F}^{-1}D\bar{p}\bar{F} \]

3. Extract and scale solutions
   \[ \tilde{v} \leftarrow \text{UniqEig}(\tilde{M}) \]
   \[ \tilde{v} \leftarrow \bar{F}^{-1}(\bar{P}^{1/2} \odot \bar{F}v) \]

4. Scale by the first moment
   \[ \rho \leftarrow C_{-1} \mu \]

\[ \bar{M} \approx CxD\rho CTx \]
The Spectral Algorithm in Full

1. Estimate moments and power spectrum

   1. $\mu \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j$
   2. $M \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j y_j^T - \sigma^2 I,$
      where $M \approx C_x D_\rho C_x^T$
   3. $P_x \leftarrow \frac{1}{N} \sum_{j=1}^{N} |Fy_j|^2 - \sigma^2 L \mathbf{1}$

   Normalize the signal

   2. $p \leftarrow (P_x)^{-1/2}$
   3. $Q \leftarrow F^{-1} D p F$
   4. $\tilde{M} \leftarrow Q M Q^{-1}$
      where $\tilde{M} \approx C_x' D_\rho C_x'^T$

   Extract and scale solutions

   1. $v \leftarrow \text{UniqEig}(\tilde{M})$
   2. $\tilde{v} \leftarrow F^{-1} \left( (P_x)^{1/2} \odot F v \right)$
   3. $x \leftarrow \left( \frac{\sum \mu}{\sum \tilde{v}} \right) \tilde{v}$
   4. $\rho \leftarrow C^{-1} x \mu$
The Spectral Algorithm in Full

1. **Estimate moments and power spectrum**
   - \( \mu \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j \)
   - \( M \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j y_j^T - \sigma^2 I \), where \( M \approx C_x D_\rho C_x^T \)
   - \( P_x \leftarrow \frac{1}{N} \sum_{j=1}^{N} |Fy_j|^2 - \sigma^2 L1 \)

2. **Normalize the signal**
   - \( p \leftarrow (P_x)^{-1/2} \)
   - \( Q \leftarrow F^{-1} D_\rho F \)
   - \( \tilde{M} \leftarrow Q M Q^{-1} \) Here \( \tilde{M} \approx C_{x'} D_\rho C_{x'}^T \) with \( C_{x'} = C_x^T \)
The Spectral Algorithm in Full

1. **Estimate moments and power spectrum**
   - \( \mu \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j \)
   - \( M \leftarrow \frac{1}{N} \sum_{j=1}^{N} y_j y_j^T - \sigma^2 I \), where \( M \approx C_x D_{\rho} C_x^T \)
   - \( P_x \leftarrow \frac{1}{N} \sum_{j=1}^{N} |Fy_j|^2 - \sigma^2 L1 \)

2. **Normalize the signal**
   - \( p \leftarrow (P_x)^{-1/2} \)
   - \( Q \leftarrow F^{-1} D_p F \)
   - \( \tilde{M} \leftarrow Q M Q^{-1} \) \quad \text{Here} \quad \tilde{M} \approx C_x' D_{\rho} C_x'^T \text{ with } C_x' = C_x^T \)

3. **Extract and scale solutions**
   - \( v \leftarrow \text{UniqEig}(\tilde{M}) \)
   - \( \tilde{v} \leftarrow F^{-1} \left( (P_x)^{1/2} \odot Fv \right) \) \quad \text{Reset the Fourier modulus}
   - \( x \leftarrow \left( \text{Sum}(\mu)/\text{Sum}(\tilde{v}) \right) \tilde{v} \) \quad \text{Scale by the first moment}
   - \( \rho \leftarrow C_x^{-1} \mu \)
Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \ldots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \ldots$. For each $n$, we draw observations $y_1, \ldots, y_{N_n}$ at noise level $\sigma_n$. 
Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \ldots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \ldots$. For each $n$, we draw observations $y_1, \ldots, y_{N_n}$ at noise level $\sigma_n$.

**Theorem**

Let $\rho$ be periodic. Then, for sufficiently small $t > 0$:

$$
P \left[ \min_s \| R_s \hat{x} - x \| \geq t \right] \leq C_1 \exp \left\{ -C_2 \frac{N_n}{\sigma_n^4} t \right\},$$

where $C_1 = C_1(x, \rho, L)$ and $C_2 = C_2(x, \rho, L)$ are finite, positive constants.
Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \ldots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \ldots$. For each $n$, we draw observations $y_1, \ldots, y_{N_n}$ at noise level $\sigma_n$.

**Theorem**

Let $\rho$ be periodic. Then, for sufficiently small $t > 0$:

$$
\mathbb{P} \left[ \min_s \| R_s \hat{x} - x \| \geq t \right] \leq C_1 \exp \left\{ -C_2 \frac{N_n}{\sigma_n^4} t \right\},
$$

where $C_1 = C_1(x, \rho, L)$ and $C_2 = C_2(x, \rho, L)$ are finite, positive constants.

Therefore, if $N_n \geq K \log(n)\sigma_n^4$ for a sufficiently large constant $K$, then the error of $\hat{x}_n$ converges to 0 almost surely as $n \to \infty$. 
The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?
The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?

**Proposition**

Let $\ell < L/2$ be a divisor of $L > 1$. Suppose that $\rho$ is a periodic distribution with period of $\ell$. Then, for a given real signal $x_1$ with non-vanishing DFT, there exists a different real signal $x_2$ (which is not a translation of $x_1$) such that both signals have the same first and second moments.
The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?

Proposition

Let $\ell < L/2$ be a divisor of $L > 1$. Suppose that $\rho$ is a periodic distribution with period of $\ell$. Then, for a given real signal $x_1$ with non-vanishing DFT, there exists a different real signal $x_2$ (which is not a translation of $x_1$) such that both signals have the same first and second moments.

A visual example:
Lower Bound

Let $\phi_x(\hat{x}) := \arg\min_{z \in \{R_\ell \hat{x}\}} \| z - x \|$ and define Minimal Square Error (MSE) to be

$$\text{MSE} = \frac{1}{\|x\|^2} \mathbb{E} \left[ \| \phi_x(\hat{x}) - x \|^2 \right].$$
Let $\phi_x(\hat{x}) := \arg\min_{z \in \{R_\ell\hat{x}\}_{\ell \in \mathbb{Z}_L}} \|z - x\|$ and define Minimal Square Error (MSE) to be

$$\text{MSE} = \frac{1}{\|x\|^2} \mathbb{E} \left[ \|\phi_x(\hat{x}) - x\|^2 \right].$$

**Theorem**

Assume that $x$ is not a constant vector. If $\hat{x}$ is an asymptotically unbiased estimator of $x$, that is $\phi_x(\hat{x}) \to x$, then

$$\text{MSE} \geq O \left( \frac{1}{\text{SNR}^2} \right).$$

Moreover, if $\rho$ is periodic, with a period $\ell < \frac{L}{2}$, then

$$\text{MSE} \geq O \left( \frac{1}{\text{SNR}^3} \right).$$
Dependency on $L$ via Spike Covariance Model

In the spiked model, given a rank $r$ matrix $X$, we observe

$$Y = X + G \in \mathbb{R}^{L \times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$. 

Let $\lambda$ be the top eigenvalue of $XX^T/N$. Then, the phase transition at $\lambda_{\text{critical}} = \sigma^2 \sqrt{\gamma}$. 

This means that for $\lambda > \lambda_{\text{critical}}$ there is a non-trivial correlation between top eigenvectors of $YY^T/N$ and $XX^T/N$. 

In MRA it is equal to $N \geq L \sigma^4 \|x\|_4^4 (\max \rho)^2 = L (\max \rho)^2 \text{SNR}^2$. 

Nir Sharon (PACM, Princeton University)
In the spiked model, given a rank $r$ matrix $X$, we observe

$$Y = X + G \in \mathbb{R}^{L \times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$. Let $\lambda$ be the top eigenvalue of $XX^T/N$. Then, the phase transition at

$$\lambda_{\text{critical}} = \sigma^2 \sqrt{\gamma}.$$

means that for $\lambda > \lambda_{\text{critical}}$ there is a non-trivial correlation between top eigenvectors of $YY^T/N$ and $XX^T/N$. 
Dependency on $L$ via Spike Covariance Model

In the spiked model, given a rank $r$ matrix $X$, we observe

\[ Y = X + G \in \mathbb{R}^{L \times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \]

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$. Let $\lambda$ be the top eigenvalue of $XX^T/N$. Then, the phase transition at

\[ \lambda_{critical} = \sigma^2 \sqrt{\gamma}. \]

means that for $\lambda > \lambda_{critical}$ there is a non-trivial correlation between top eigenvectors of $YY^T/N$ and $XX^T/N$.

In MRA it is equal to

\[ N \geq \frac{L\sigma^4}{\|x\|^4 (\max \rho)^2} = \frac{L}{(\max \rho)^2} \frac{1}{\text{SNR}^2}. \]
Table of contents

1 The Problem of MultiReference Alignment (MRA)

2 Moments Approach and Theoretical Guarantees

3 Numerical Examples and Conclusions
Two additional algorithms – LS optimization

The least square algorithm aim to minimize

$$\min_{\tilde{x} \in \mathbb{R}^L, \tilde{\rho} \in \Delta^L} \| \hat{M}y - C\tilde{x}D\tilde{\rho}C^T \|_F^2 + \lambda \| \hat{\mu}y - C\tilde{x}\tilde{\rho} \|_2^2,$$

where $\lambda > 0$ is a predefined parameter.

In low SNR regime, the variance of the first estimator is proportional to $\sigma^2$ and the variance of the second is proportional to $3L\sigma^4$. Therefore, we set $\lambda = \frac{1}{L(1+3\sigma^2)}$.

We solve the problem with a gradient-based method.
Two additional algorithms – adapted EM

By the model of the MRA,

\[ p(y, \ell|x, \rho) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \left\| R_{r_j} x - y_{j} \right\|^2} \rho[j]. \]

The log-likelihood function is then given, up to a constant, by

\[ \log L(x, \rho|y, r) = \sum_{j=1}^{N} \left\{ \log \rho[j] - \frac{1}{2\sigma^2} \left\| R_{r_j} x - y_{j} \right\|^2 \right\}. \]

To maximize the expectation of the marginal log-likelihood we use the iteration

\[ x_{k+1} = \frac{1}{N} \sum_{j=1}^{N} \sum_{\ell=0}^{L-1} w_{k}^{\ell,j} R_{\ell}^{-1} y_{j} \quad \text{and} \quad \rho_{k+1}[\ell] = \frac{1}{N} \sum_{k=1}^{N} \sum_{\ell'=0}^{L-1} \frac{w_{k}^{\ell,j}}{w_{k}^{\ell',j}}. \]

for weights which are iteratively updated according to the data.

\[ w_{k}^{\ell,j} = \mathbb{P}[r_j = \ell|y, x_k, \rho_k]. \]
EM: Standard vs Adapted

A family of distributions

The uniformity factor $s$

Relative error

Nir Sharon (PACM, Princeton University)
Multireference Alignment
October 11, 2017 20 / 24
EM: Standard vs Adapted

A family of distributions

![Graph showing A family of distributions with curves for s = 3, s = 4.75, and s = 7.](image)
EM: Standard vs Adapted

A family of distributions

A comparison of standard EM and our adapted version for non-uniform distributions,
Comparing the Different Methods

The relative error as a function of varying level of noise, $\sigma$. 
The relative error as a function of varying level of noise, $\sigma$. 

![Graph comparing the relative error of different methods as a function of noise level. The x-axis represents noise level ($\sigma$) on a logarithmic scale, and the y-axis represents relative error on a logarithmic scale. Three curves are shown for LS (dashed blue), Adapted EM (solid red), and Spectral (dotted black). The graph illustrates how each method performs under varying noise conditions.]
We expect the (unsquared) error to be

\[
\min_{s \in (\mathbb{Z})^L} \| R_s \hat{x} - x \| \leq \frac{\sigma^d}{\sqrt{N}},
\]

where \( d = 1 \) for low level of noise (via alignment) whereas \( d = 2 \) for high level of noise (as implied by the spectral algorithm).
We expect the (unsquared) error to be

\[
\min_{s \in (\mathbb{Z})^L} \| R_s \hat{x} - x \| \leq \frac{\sigma^d}{\sqrt{N}},
\]

where \( d = 1 \) for low level of noise (via alignment) whereas \( d = 2 \) for high level of noise (as implied by the spectral algorithm).

Numerical evidence to the error rates: least squares fitting in a log-log plot.
Concluding Remarks

- We can solve the MRA problem in regimes of high level of noise by using low order statistics and simultaneously targeting the signal and the distribution.

- The sample complexity of the problem is $N \gtrsim \sigma^4$ for aperiodic distributions and $N \gtrsim \sigma^6$ for periodic distributions. That makes the aperiodic case easier!

- Numerical examples confirm the theory and show the advantage one gets of including the distribution into the estimator model.
Th-th-th-th-th-that’s all folks!

Thank you

Reference: