Estimation in extreme noise levels with application to cryo-electron microscopy

Tamir Bendory

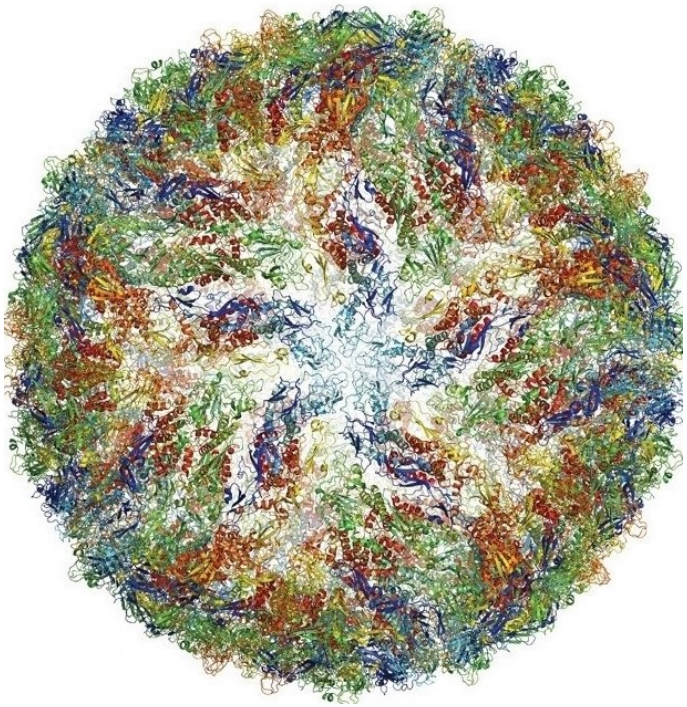
December 12, 2018

Princeton University
The Program in Applied and Computational Mathematics
https://web.math.princeton.edu/~tdory

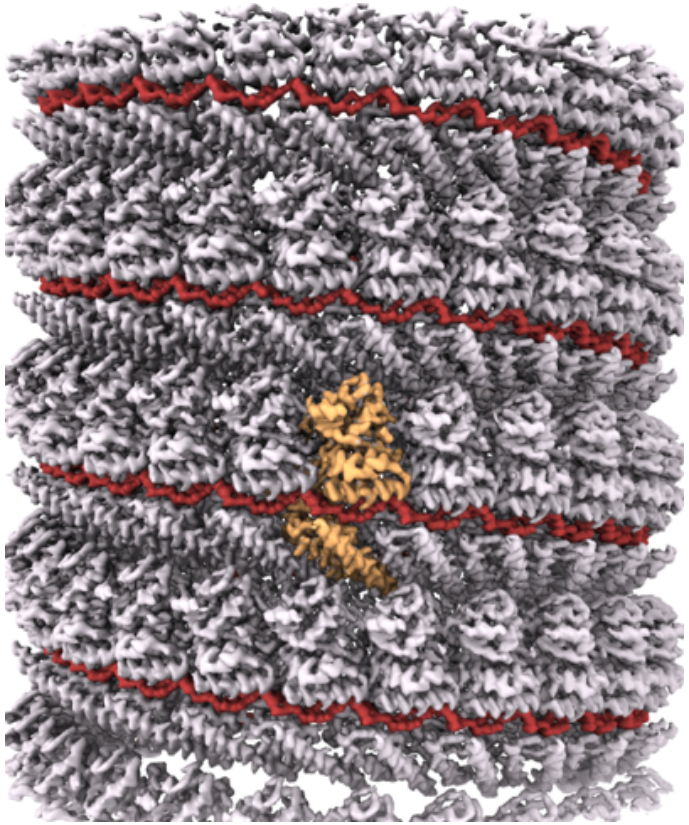
Outline

Outline

The resolution revolution



The Zika virus



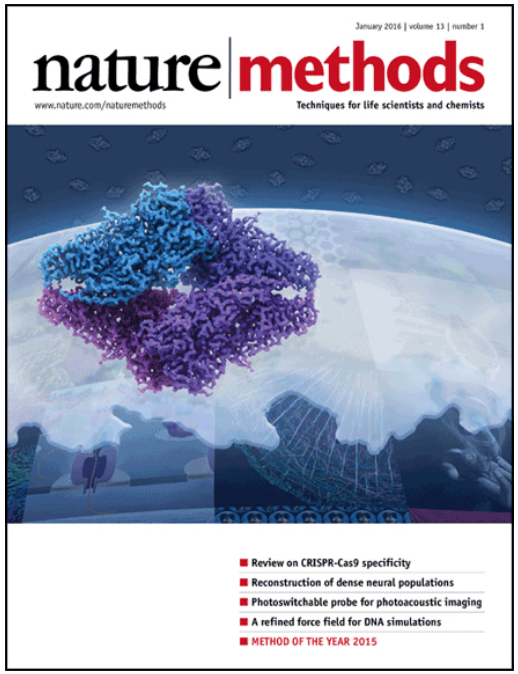
The Ebola virus

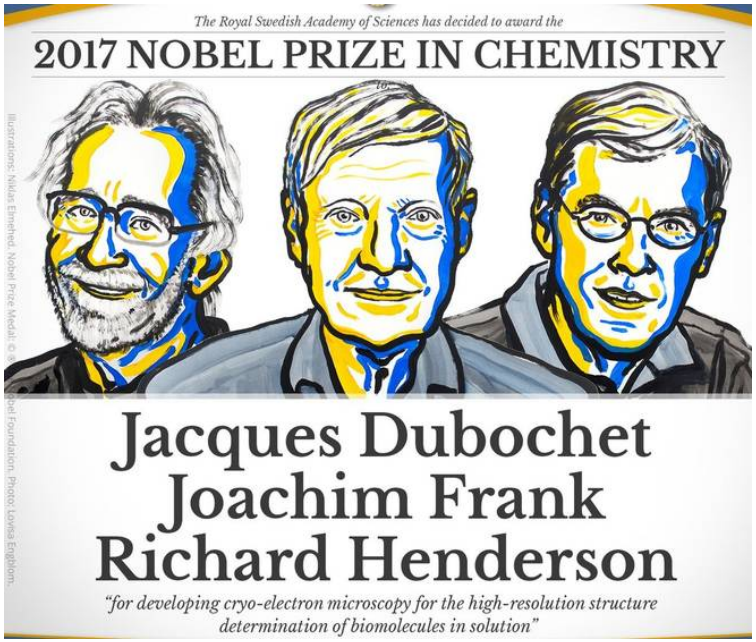
In biology, a key idea is that **structure determines function**

---

https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2017/; Sugita et al., 2018

Exciting times for cryo-EM





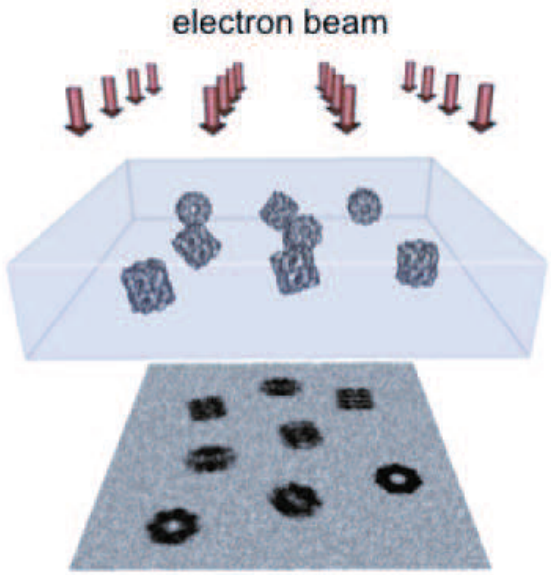
Method of the Year 2015

"Single-particle cryo-EM is our choice for Method of the Year 2015 for its newfound ability to solve protein structures at near-atomic resolution."

Nobel Prize in Chemistry 2017

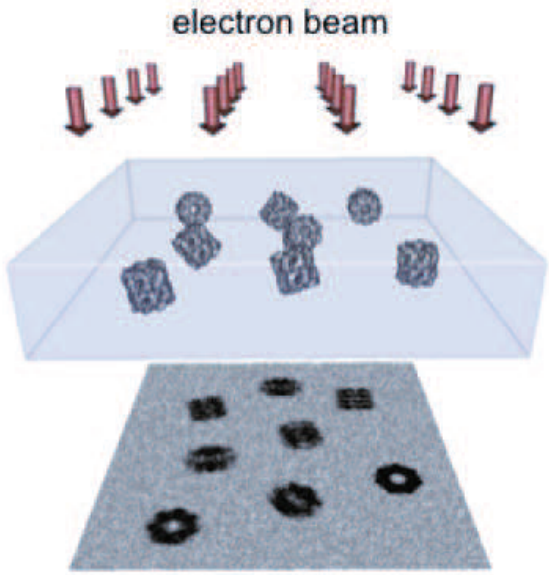
"for developing cryo–EM for the high-resolution structure determination of biomolecules in solution"

How does it work?

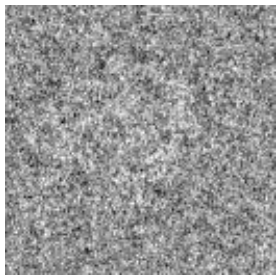
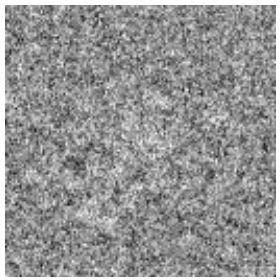
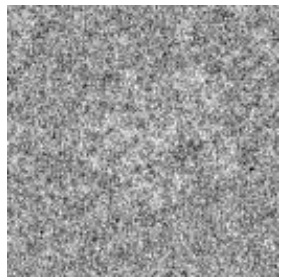
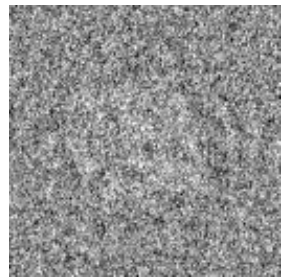


The images of E. coli 50S ribosomal subunit were provided by Dr. Fred Sigworth, Yale Medical School.

How does it work?

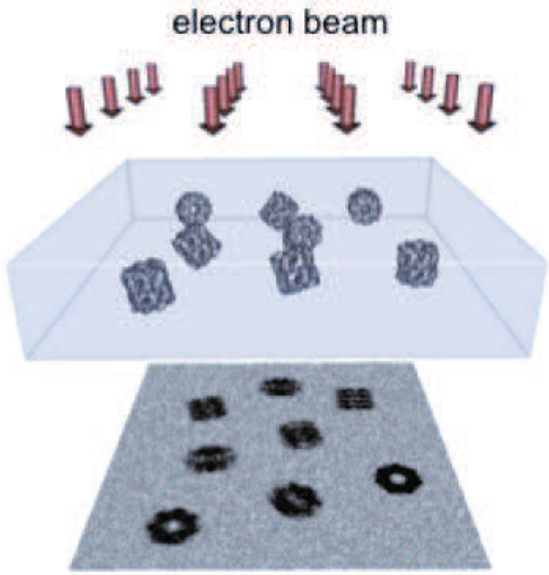


- High noise level

The images of E. coli 50S ribosomal subunit were provided by Dr. Fred Sigworth, Yale Medical School.

How does it work?



- High noise level
- Parameters to be estimated vs. nuisance variables
  - The goal is to estimate the 3-D structure
  - All other unknowns are **nuisance variables**

The images of E. coli 50S ribosomal subunit were provided by Dr. Fred Sigworth, Yale Medical School.

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

- consistent

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

- consistent
- asymptotically efficient

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

- consistent
- asymptotically efficient
- hard to compute (usually, involves solving a non-convex, high-dimensional problem)

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

- consistent
- asymptotically efficient
- hard to compute (usually, involves solving a non-convex, high-dimensional problem)

Parameter estimation

The most popular techniques in cryo-EM are based on **maximum likelihood estimation**:

- consistent
- asymptotically efficient
- hard to compute (usually, involves solving a non-convex, high-dimensional problem)

Alternative: **the method of moments** (Pearson, 1894)

Method of moments

Suppose that the distribution of y is parametrized by x. The goal is to **estimate x from observations of y.**

Method of moments

Suppose that the distribution of y is parametrized by x. The goal is to **estimate x from observations of y**.

Steps:

1. Derive the population moments
2. Estimate the moments from the data (sample moments)
3. Solve the (polynomial) system of equations

$$Ey = p_1(x)$$
$$Eyy^T = p_2(x)$$
$$\vdots$$

Method of moments

Suppose that the distribution of y is parametrized by x. The goal is to **estimate x from observations of y**.

Steps:

1. Derive the population moments
2. Estimate the moments from the data (sample moments)
3. Solve the (polynomial) system of equations

$$M_1 := \frac{1}{N}\sum_{i=1}^{N} y_i \approx Ey = p_1(x)$$

$$M_2 := \frac{1}{N}\sum_{i=1}^{N} y_i y_i^T \approx Eyy^T = p_2(x)$$

$$\vdots$$

Method of moments

Suppose that the distribution of y is parametrized by x. The goal is to **estimate x from observations of y**.

Steps:

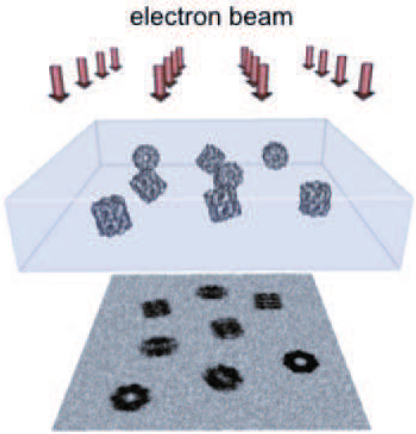
1. Derive the population moments
2. Estimate the moments from the data (sample moments)
3. Solve the (polynomial) system of equations

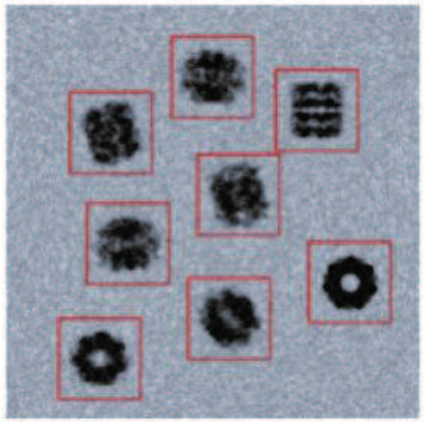
$$M_1 := \frac{1}{N}\sum_{i=1}^{N} y_i \approx Ey = p_1(x)$$

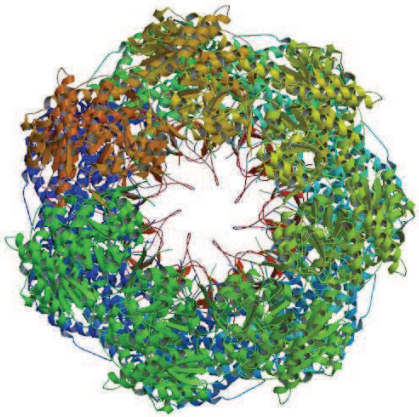
$$M_2 := \frac{1}{N}\sum_{i=1}^{N} y_i y_i^T \approx Eyy^T = p_2(x)$$

$$\vdots$$

The cryo-EM inverse problem

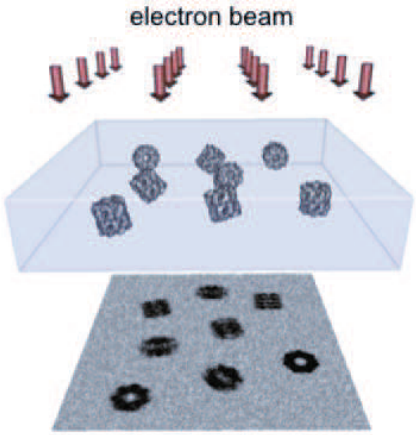


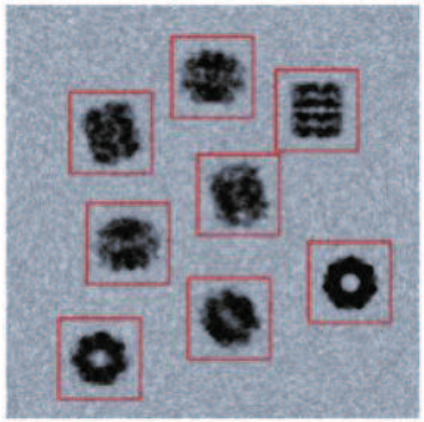


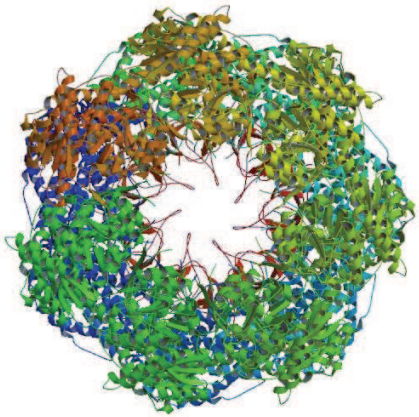


micrographs (data) → particle picking → reconstruction algorithm

The cryo-EM inverse problem







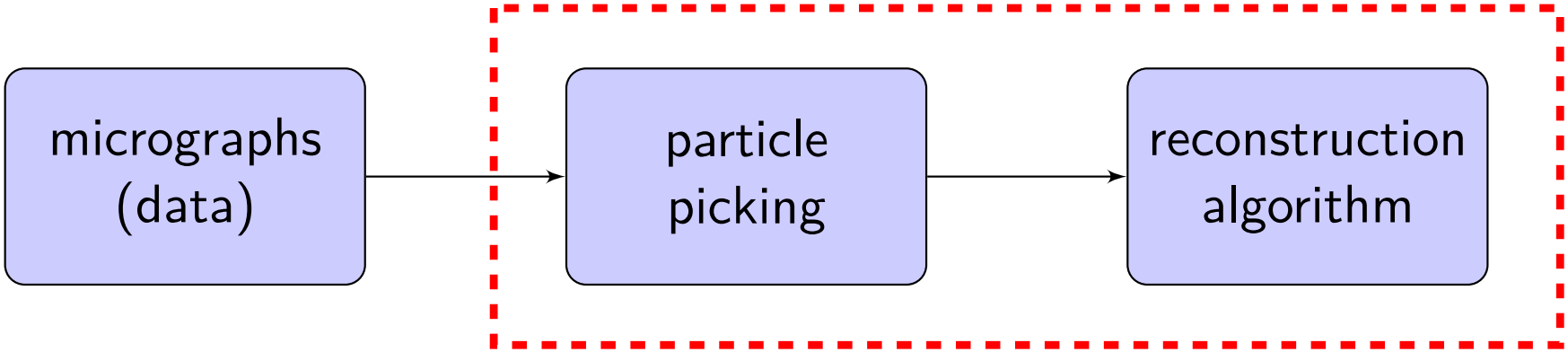


Image formation model (perfect particle picking)

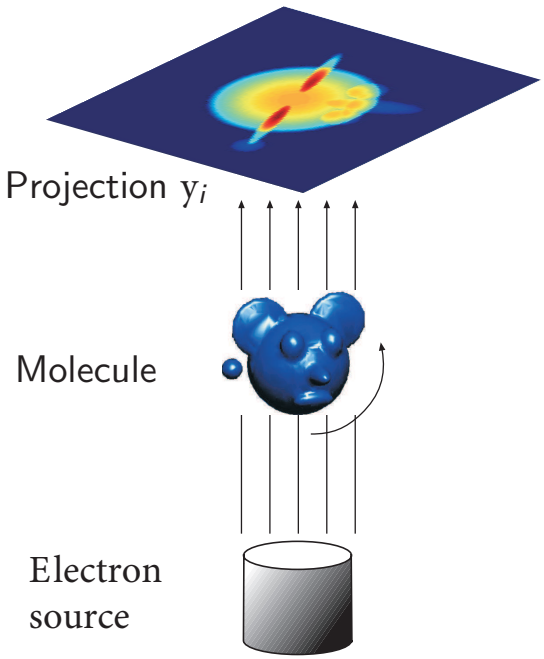


Image formation model

$$y_i = P(R_i \circ x) + \text{noise}, \; R_i \in SO(3)$$

Image formation model (perfect particle picking)

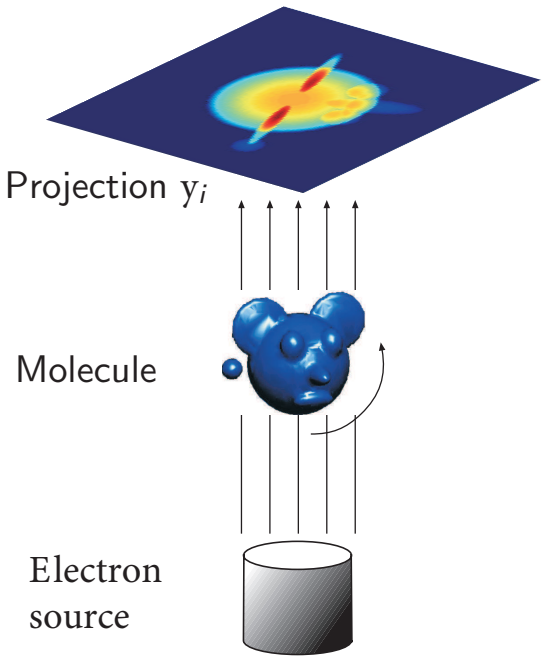


Image formation model

$$y_i = P(R_i \circ x) + \text{noise}, \; R_i \in SO(3)$$

The cryo–EM problem

Estimate x given $y_1, \ldots, y_N$

Image formation model (perfect particle picking)

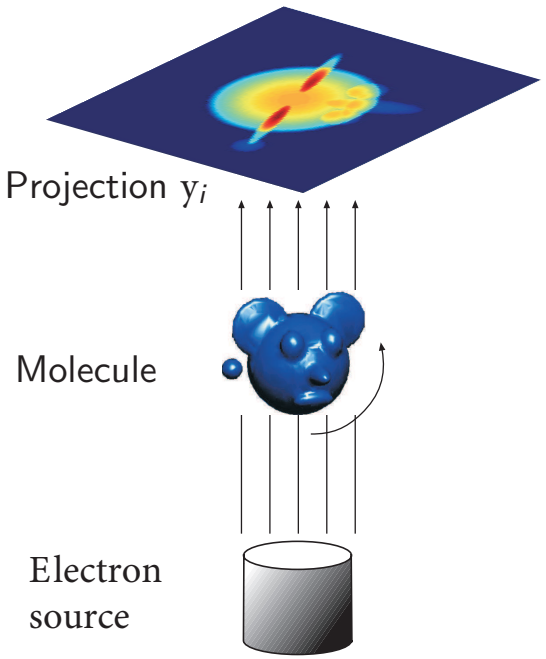


Image formation model

$$y_i = P(R_i \circ x) + \text{noise}, \; R_i \in SO(3)$$

The cryo–EM problem

Estimate x given $y_1, \ldots, y_N$

- Can we accurately estimate the rotations?
- Can we accurately estimate the volume x?
- And how?
- What is the optimal estimation rate?

The multi-reference alignment (MRA) problem

Problem: Estimate a signal $x \in \mathbb{R}^L$, up to cyclic shift, from its noisy circularly-shifted copies

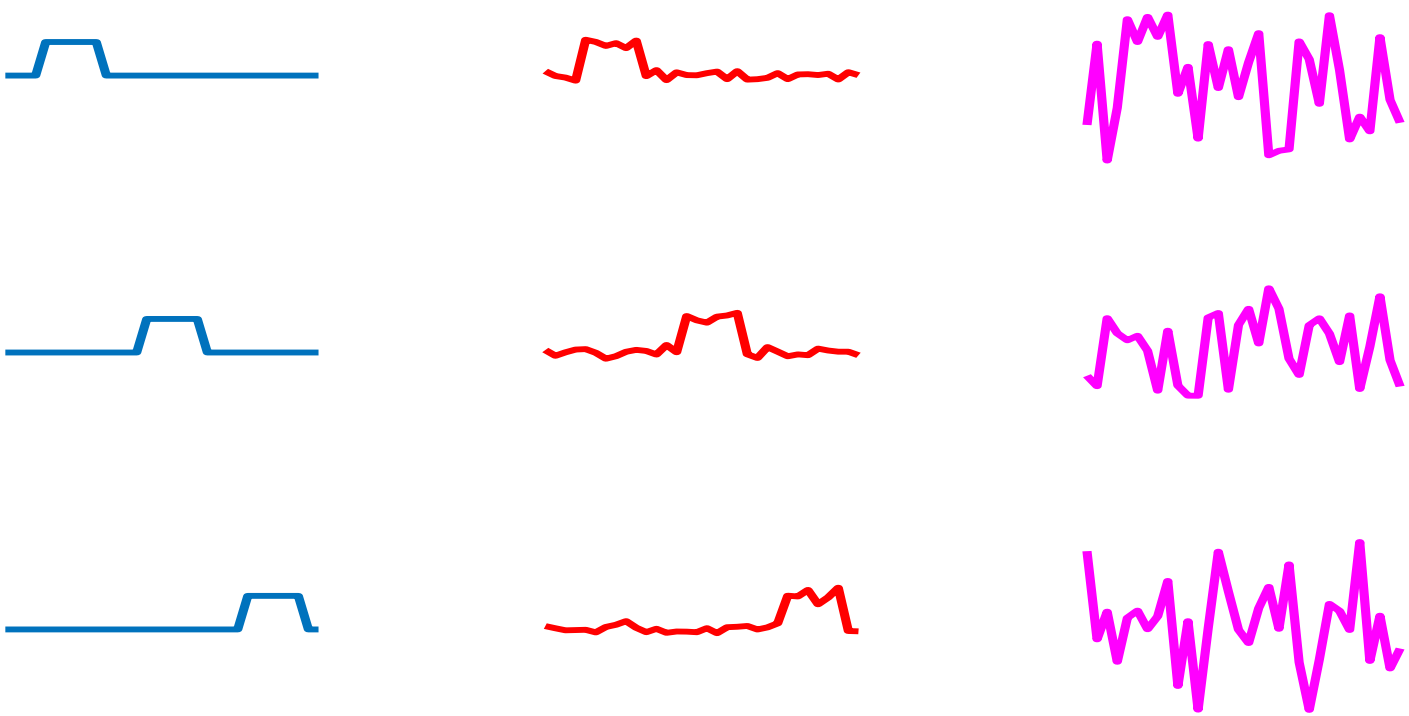
$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

The multi-reference alignment (MRA) problem

Problem: Estimate a signal $x \in \mathbb{R}^L$, up to cyclic shift, from its noisy circularly-shifted copies

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

$\sigma = 0$ $\sigma = 0.1$ $\sigma = 1.2$



The multi-reference alignment (MRA) problem

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Given the shifts, we can estimate

$$\frac{1}{N} \sum_{i=1}^{N} R_{r_i}^{-1} y_i \to x$$

Unbiased estimator, variance σ^2/N

The multi-reference alignment (MRA) problem

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Given the shifts, we can estimate

$$\frac{1}{N} \sum_{i=1}^{N} R_{r_i}^{-1} y_i \to x$$

Unbiased estimator, variance σ^2/N

Can we estimate the shifts?

Alignment is impossible in low SNR

- Let

$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$

Alignment is impossible in low SNR

- Let

$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$

- Let

$$y = \begin{cases} R_1 x + \varepsilon, & \text{if } \eta = 0, \\ R_2 x + \varepsilon, & \text{if } \eta = 1, \end{cases}$$

where $\varepsilon \sim N(0, \sigma^2 I)$.

Alignment is impossible in low SNR

- Let
$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$
- Let
$$y = \begin{cases} R_1 x + \varepsilon, & \text{if } \eta = 0, \\ R_2 x + \varepsilon, & \text{if } \eta = 1, \end{cases}$$
where $\varepsilon \sim N(0, \sigma^2 I)$.
- We observe y. How reliably can we estimate η?

Alignment is impossible in low SNR

- Let
$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$
- Let
$$y = \begin{cases} R_1 x + \varepsilon, & \text{if } \eta = 0, \\ R_2 x + \varepsilon, & \text{if } \eta = 1, \end{cases}$$
where $\varepsilon \sim N(0, \sigma^2 I)$.
- We observe y. How reliably can we estimate η?
- The estimator $\hat{\eta} = 0$ is correct with probability $1/2$. Can we do better?

Alignment is impossible in low SNR

- Let
$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$
- Let
$$y = \begin{cases} R_1 x + \varepsilon, & \text{if } \eta = 0, \\ R_2 x + \varepsilon, & \text{if } \eta = 1, \end{cases}$$
where $\varepsilon \sim N(0, \sigma^2 I)$.
- We observe y. How reliably can we estimate η?
- The estimator $\hat{\eta} = 0$ is correct with probability $1/2$. Can we do better?

Alignment is impossible in low SNR

- Let
$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$
- Let
$$y = \begin{cases} R_1 x + \varepsilon, & \text{if } \eta = 0, \\ R_2 x + \varepsilon, & \text{if } \eta = 1, \end{cases}$$
where $\varepsilon \sim N(0, \sigma^2 I)$.
- We observe y. How reliably can we estimate η?
- The estimator $\hat{\eta} = 0$ is correct with probability $1/2$. Can we do better?

Proposition (Bendory, Boumal, Leeb, Levin, Singer, '18)

For any deterministic estimator $\hat{\eta}$ of η,

$$\lim_{\sigma \to \infty} Prob(\hat{\eta} = \eta) \leq 1/2.$$

Alignment is impossible in low SNR

- Let
$$\eta = \begin{cases} 0, & \text{with probablity } 1/2, \\ 1, & \text{with probablity } 1/2. \end{cases}$$
- Let
$$y = \begin{cases} x_1 + \varepsilon, & \text{if } \eta = 0, \\ x_2 + \varepsilon, & \text{if } \eta = 1, \end{cases}$$
where $\varepsilon \sim N(0, \sigma^2 I)$.
- We observe y. How reliably can we estimate η?
- The estimator $\hat{\eta} = 0$ is correct with probability $1/2$. Can we do better?

Proposition (Bendory, Boumal, Leeb, Levin, Singer, '18)

For any deterministic estimator $\hat{\eta}$ of η,

$$\lim_{\sigma \to \infty} Prob(\hat{\eta} = \eta) \leq 1/2.$$

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.
- Assign probabilities (E-step): $\omega_i^\ell \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_{\text{est}} - y_i\|_2^2\right)$

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.
- Assign probabilities (E-step): $\omega_i^\ell \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_{\text{est}} - y_i\|_2^2\right)$
- Update estimate (M-step): $x_{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_i^\ell R_\ell^{-1} y_i$

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.
- Assign probabilities (E-step): $\omega_i^\ell \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_{\text{est}} - y_i\|_2^2\right)$
- Update estimate (M-step): $x_{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_i^\ell R_\ell^{-1} y_i$
- Works well numerically, but may converge to local optimum (difficult to analyze mathematically)

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.
- Assign probabilities (E-step): $\omega_i^\ell \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_{\text{est}} - y_i\|_2^2\right)$
- Update estimate (M-step): $x_{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_i^\ell R_\ell^{-1} y_i$
- Works well numerically, but may converge to local optimum (difficult to analyze mathematically)
- Slow in a low SNR environment

Expectation-maximization for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- Expectation-maximization (EM) is an iterative method to find the marginalized maximum likelihood.
- Assign probabilities (E-step): $\omega_i^\ell \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_{\text{est}} - y_i\|_2^2\right)$
- Update estimate (M-step): $x_{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_i^\ell R_\ell^{-1} y_i$
- Works well numerically, but may converge to local optimum (difficult to analyze mathematically)
- Slow in a low SNR environment
- Can we achieve similar performance with **one pass over the data**?

Method of moments for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- First moment:

$$Ey = x * \rho,$$

where ρ is the distribution of the shifts

Method of moments for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- First moment:

$$Ey = x * \rho,$$

  where ρ is the distribution of the shifts

- Second moment: (up to constant bias term)

$$Eyy^T = C_x D_\rho C_x^T,$$

  where C_x is a circulant matrix of x and D_ρ is a diagonal matrix of ρ

Method of moments for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- First moment:

$$Ey = x * \rho,$$

  where ρ is the distribution of the shifts

- Second moment: (up to constant bias term)

$$Eyy^T = C_x D_\rho C_x^T,$$

  where C_x is a circulant matrix of x and D_ρ is a diagonal matrix of ρ

- Individual shifts do not appear in the moments.

Method of moments for MRA

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \dots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

- First moment:

$$Ey = x * \rho,$$

  where ρ is the distribution of the shifts

- Second moment: (up to constant bias term)

$$Eyy^T = C_x D_\rho C_x^T,$$

  where C_x is a circulant matrix of x and D_ρ is a diagonal matrix of ρ

- Individual shifts do not appear in the moments.

- **Is the second moment enough?**

Spectral algorithm

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Algorithm:

1. Compute $Eyy^T = C_x D_\rho C_x^T$

Spectral algorithm

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Algorithm:

1. Compute $Eyy^T = C_x D_\rho C_x^T$
2. Rotate $z_i = R_{s_i} y_i$, where s_i are drawn from the uniform distribution

Spectral algorithm

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Algorithm:

1. Compute $Eyy^T = C_x D_\rho C_x^T$
2. Rotate $z_i = R_{s_i} y_i$, where s_i are drawn from the uniform distribution
3. Compute $Ezz^T = \frac{1}{L} C_x C_x^T$

Spectral algorithm

Model:

$$y_i = R_{r_i} x + \varepsilon_i, \quad i = 1, \ldots, N, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

Algorithm:

1. Compute $Eyy^T = C_x D_\rho C_x^T$
2. Rotate $z_i = R_{s_i} y_i$, where s_i are drawn from the uniform distribution
3. Compute $Ezz^T = \frac{1}{L} C_x C_x^T$
4. Compute the eigendecomposition of

$$M = L \left(Eyy^T\right) \left(Ezz^T\right)^{-1} = C_x D_\rho C_x^{-1},$$

assuming C_x is invertible.

Recovery from the second moment

- We compute the eigendecomposition of

$$M = L\left(Eyy^T\right)\left(Ezz^T\right)^{-1} = C_x D_\rho C_x^{-1}$$

Recovery from the second moment

- We compute the eigendecomposition of

$$M = L\left(Eyy^T\right)\left(Ezz^T\right)^{-1} = C_x D_\rho C_x^{-1}$$

- Any eigenvector of M associated with a unique entry of ρ is a shifted, scaled version of x. The scale is fixed by the first moment.

Recovery from the second moment

- We compute the eigendecomposition of

$$M = L\left(Eyy^T\right)\left(Ezz^T\right)^{-1} = C_x D_\rho C_x^{-1}$$

- Any eigenvector of M associated with a unique entry of ρ is a shifted, scaled version of x. The scale is fixed by the first moment.

Recovery from the second moment

- We compute the eigendecomposition of

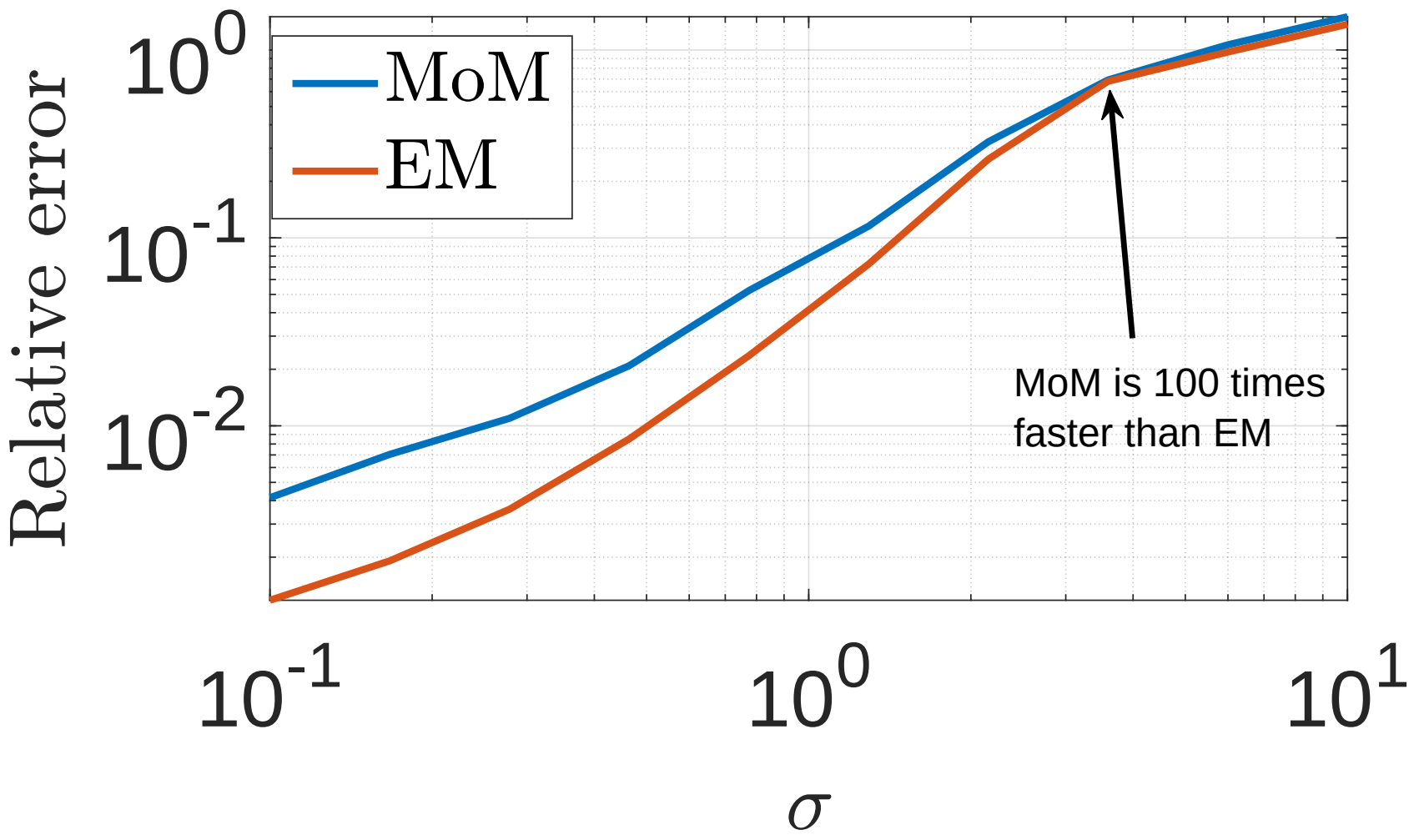
$$M = L\left(Eyy^T\right)\left(Ezz^T\right)^{-1} = C_x D_\rho C_x^{-1}$$

- Any eigenvector of M associated with a unique entry of ρ is a shifted, scaled version of x. The scale is fixed by the first moment.

Theorem (Bendory et al.,'17; Abbe et al.,'17; Ma et al.,'18)

1. *If C_x is invertible and ρ is non-periodic, then the second moment determines the signal uniquely (up to cyclic shift).*
2. *For periodic distributions (e.g., uniform), the third moment is enough.*
3. *In the low SNR regime, the method of moments achieves the optimal estimation rate.*

Numerical experiment



measurements $= 10^5$, 20 trials per point, random signal and distribution of length $= 15$

Properties of the method of moments

- Easy to compute
- Requires only one pass over the data
- Parallelizable
- Consistent (empirically)

Application to cryo-EM

Theorem (Levin, Bendory, Boumal, Kileel, Singer, '17)

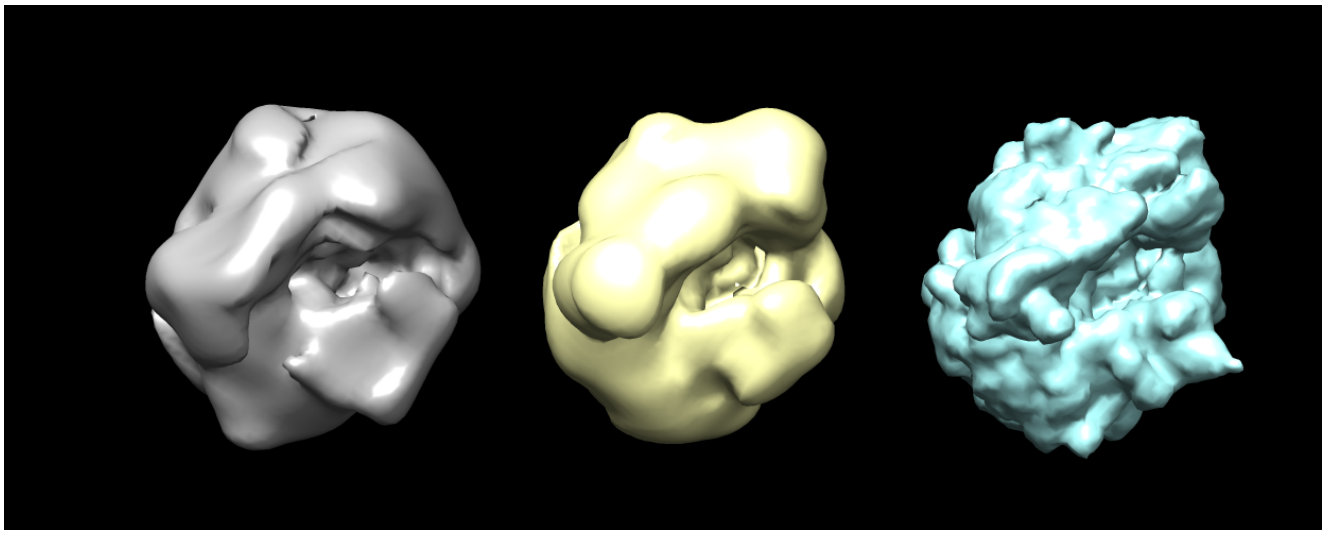
A generic volume is determined uniquely from the second moment of the projection images and two clean projections (under some conditions).

Based on (Kam, 1980)

Application to cryo-EM

Theorem (Levin, Bendory, Boumal, Kileel, Singer, '17)

A generic volume is determined uniquely from the second moment of the projection images and two clean projections (under some conditions).

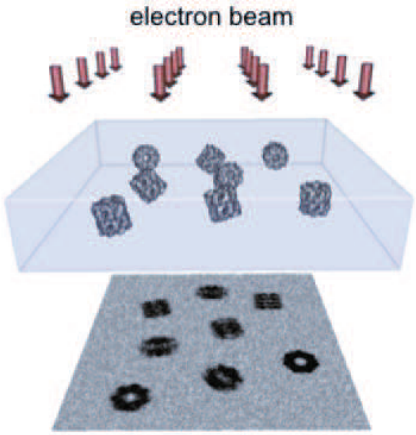


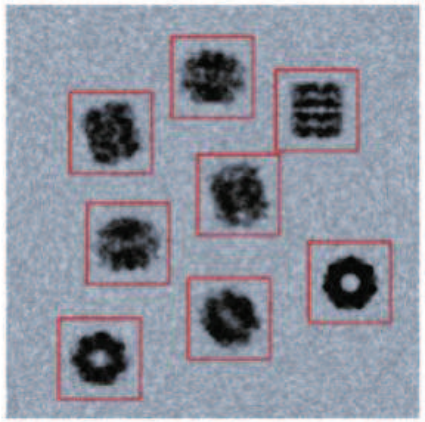
estimated structure (gray), low-resolution structure (yellow), high-resolution structure (blue)
70S ribosome with P-site tRNA, $50,000$ projections of size 109^2, $L = 10$, SNR $= 1/10$

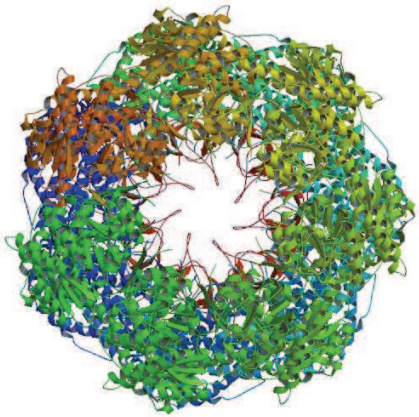
Based on (Kam, 1980)

1. Introduction
2. Multi-reference alignment
3. Estimation below the detection limit
4. Future work

The cryo-EM inverse problem







micrographs (data)

particle picking

reconstruction algorithm

estimation directly from the micrograph?

Can we estimate small molecules using cryo-EM?

Common belief: Small molecules cannot be reconstructed using cryo-EM.

For instance, see (Henderson, 1995) and (Frank, '17).

Picture credit (Heimowitz et al., '18)

Can we estimate small molecules using cryo-EM?

Common belief: Small molecules cannot be reconstructed using cryo-EM.

Reasoning:

small molecules

$\Downarrow$

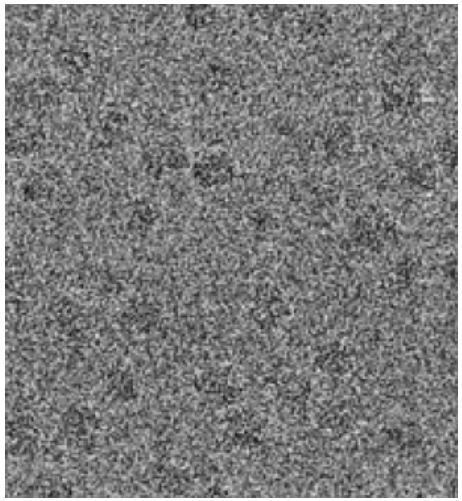
low SNR

$\Downarrow$

detection is impossible

$\Downarrow$

3-D reconstruction is impossible



For instance, see (Henderson, 1995) and (Frank, '17).

Picture credit (Heimowitz et al., '18)

Can we estimate small molecules using cryo-EM?

Common belief: Small molecules cannot be reconstructed using cryo-EM.

Reasoning:

small molecules

$\Downarrow$

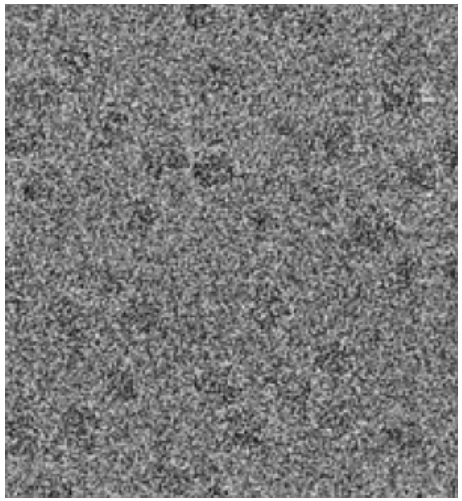
low SNR

$\Downarrow$

detection is impossible

$\Downarrow$

3-D reconstruction is impossible



Motivation: If reconstruction is possible without detection, even small molecules should be within reach for cryo-EM.

For instance, see (Henderson, 1995) and (Frank, '17).

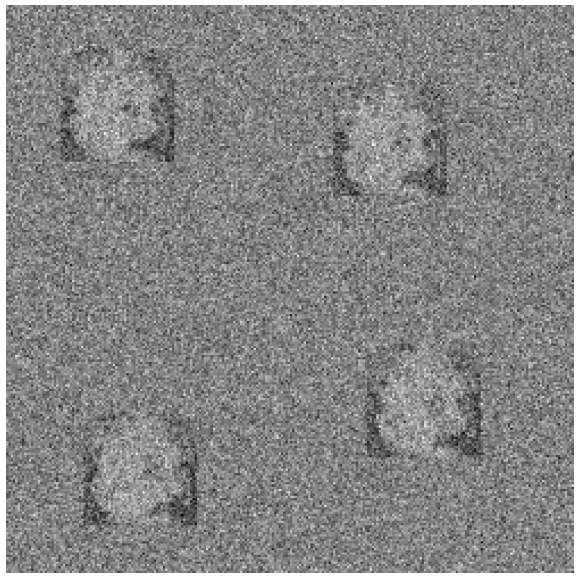
Picture credit (Heimowitz et al., '18)

Simplified model for cryo-EM (blind deconvolution)

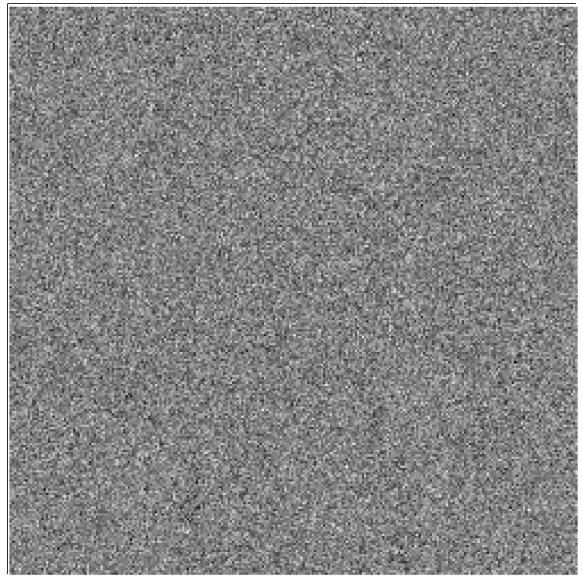
Problem: Multiple occurrences of x are embedded at random locations in a noisy measurement (micrograph) y

Goal: Estimating x from y

(a) $\sigma = 0$



(b) $\sigma = 0.5$



(c) $\sigma = 3$

Mathematical formulation

- **Problem**: Estimate $x \in \mathbb{R}^L$ from

$$y = x * s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad s \in \{0,1\}^N$$

Mathematical formulation

- **Problem**: Estimate $x \in \mathbb{R}^L$ from

$$y = x * s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad s \in \{0,1\}^N$$

- Given s (nuisance variables), estimating x is easy.

Mathematical formulation

- **Problem**: Estimate $x \in \mathbb{R}^L$ from

$$y = x * s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad s \in \{0,1\}^N$$

- Given s (nuisance variables), estimating x is easy.
- But we cannot estimate s in the low SNR regime.

Mathematical formulation

- **Problem**: Estimate $x \in \mathbb{R}^L$ from

$$y = x * s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad s \in \{0,1\}^N$$

- Given s (nuisance variables), estimating x is easy.
- But we cannot estimate s in the low SNR regime.
- **Simplifying assumption**: Any two nonzero entries of s are separated by at least $2L-1$ entries.

Mathematical formulation

- **Problem**: Estimate $x \in \mathbb{R}^L$ from

$$y = x * s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad s \in \{0,1\}^N$$

- Given s (nuisance variables), estimating x is easy.
- But we cannot estimate s in the low SNR regime.
- **Simplifying assumption**: Any two nonzero entries of s are separated by at least $2L-1$ entries.
- **Main tool**: Autocorrelation analysis

$$a_z^2[\ell] = \frac{1}{L}\sum_i z[i]z[i+\ell]$$

$$a_z^3[\ell_1,\ell_2] = \frac{1}{L}\sum_i z[i]z[i+\ell_1]z[i+\ell_2]$$

Estimating a signal from autocorrelations

In the limit $N \to \infty$:

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \text{bias}, \qquad \ell = 0, \ldots, L-1,$$
$$a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \text{bias}, \qquad \ell_1, \ell_2 = 0, \ldots, L-1,$$

where

$$\gamma := \frac{ML}{N} \leq 1.$$

Estimating a signal from autocorrelations

In the limit $N \to \infty$:

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \text{bias}, \qquad \ell = 0, \ldots, L-1,$$
$$a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \text{bias}, \qquad \ell_1, \ell_2 = 0, \ldots, L-1,$$

where

$$\gamma := \frac{ML}{N} \leq 1.$$

Theorem (Bendory, Boumal, Leeb, Levin, Singer, '18)

The signal x, the density γ and noise variance σ^2 are determined uniquely from a_y^2 and a_y^3 under mild conditions.

Estimating a signal from autocorrelations

In the limit $N \to \infty$:

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \text{bias}, \qquad \ell = 0, \ldots, L-1,$$
$$a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \text{bias}, \qquad \ell_1, \ell_2 = 0, \ldots, L-1,$$

where

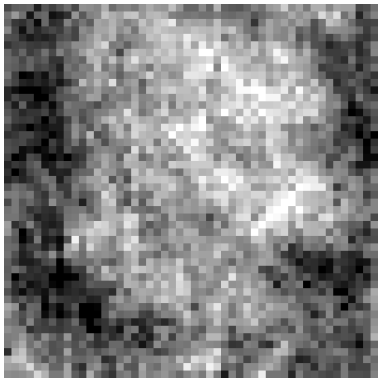
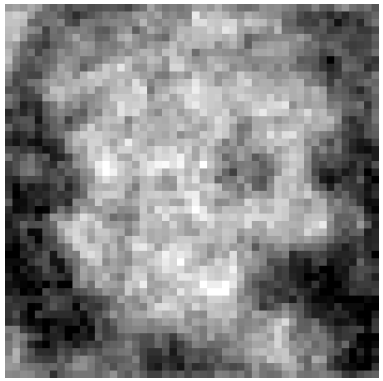
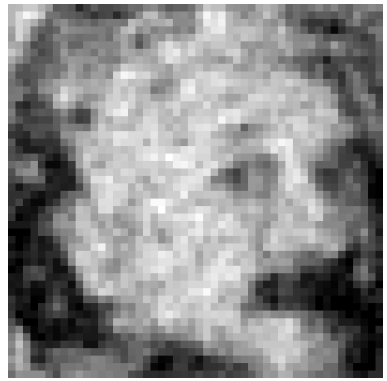
$$\gamma := \frac{ML}{N} \leq 1.$$

Theorem (Bendory, Boumal, Leeb, Levin, Singer, '18)

The signal x, the density γ and noise variance σ^2 are determined uniquely from a_y^2 and a_y^3 under mild conditions.

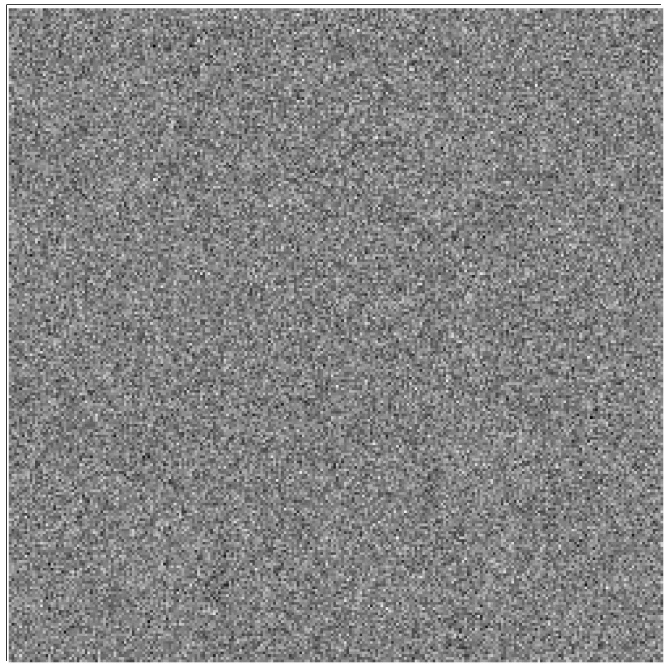
The signal x is determined, **without intermediate detection**!

Numerical experiments

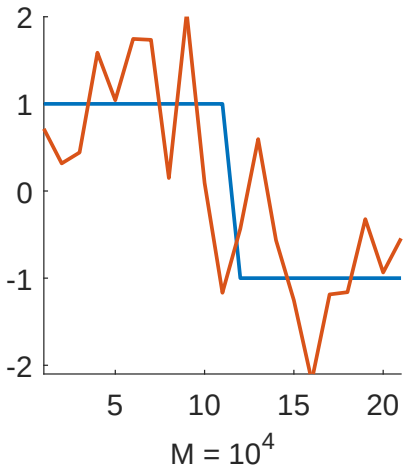




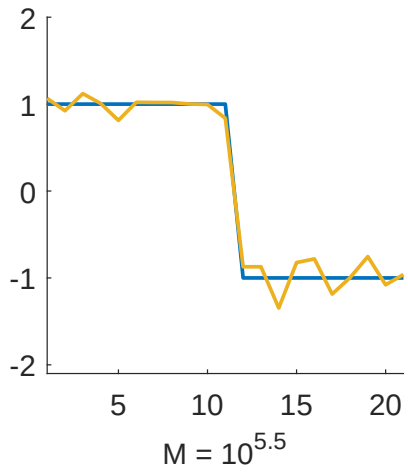

Details:

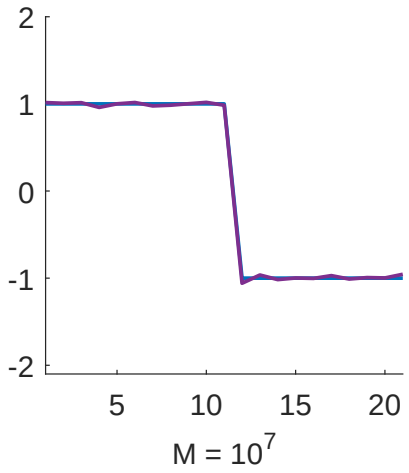
γ and σ are **known**
Recovery by relaxed-reflect-reflect (RRR)
$\sigma = 3$
Micrograph size $= 4096 \times 4096$
Image size $= 50 \times 50$
$\#$ micrographs $= 2 \cdot 10^2, 2 \cdot 10^3, 2 \cdot 10^4, 2 \cdot 10^5$
700 image occurrences on average per micrograph

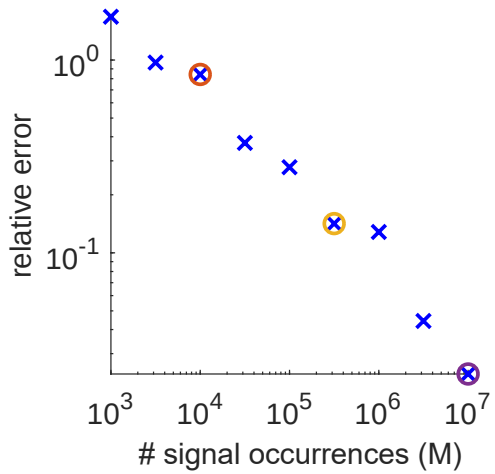


Numerical experiments



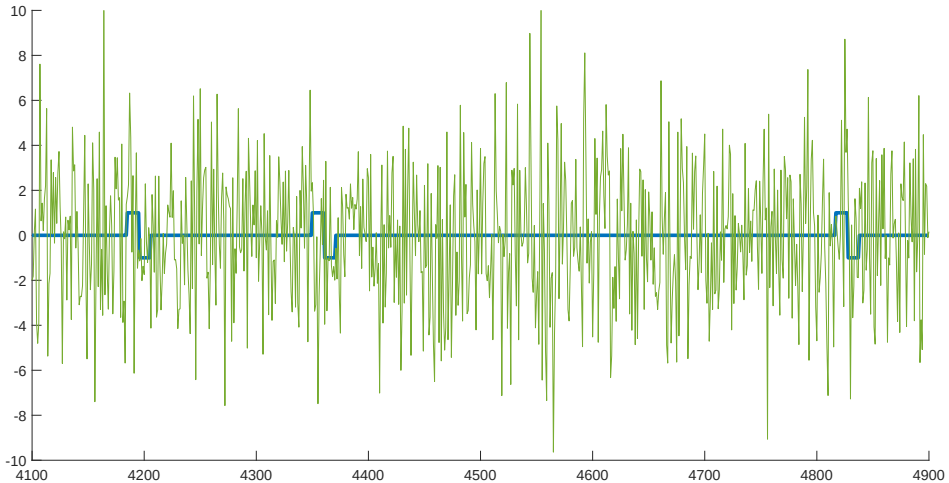






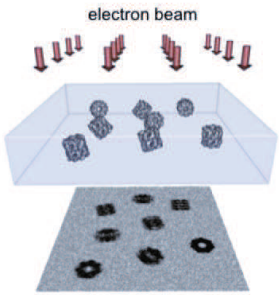
Details:

γ and σ are **unknown**
Recovery by least-squares
$\sigma = 3$
Micrograph size $= 10M(2L-1)$
Relative error $\gamma = 4.8\%, 4\%, 1.2\%$

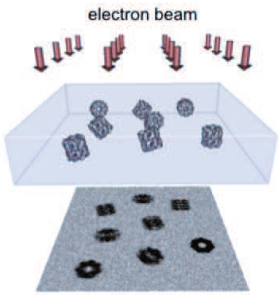


Application to cryo-EM

- In the cryo–EM setup, we aim at estimating the 3-D volume directly from the micrograph.



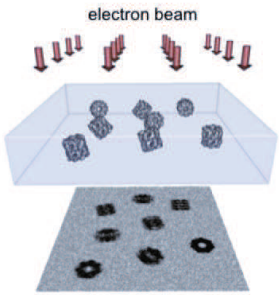
Application to cryo-EM



- In the cryo–EM setup, we aim at estimating the 3-D volume directly from the micrograph.

- An L-bandlimited 3-D volume is described by $\sim L^3$ parameters.

Application to cryo-EM



- In the cryo–EM setup, we aim at estimating the 3-D volume directly from the micrograph.

- An L-bandlimited 3-D volume is described by $\sim L^3$ parameters.

- We consider a simplified model:

  - The projections are separated
  - Gaussian noise
  - No contrast transfer function
  - Uniform distribution of viewing directions

How does it work for cryo-EM?

- We scan the micrographs with a sliding window of size $L \times L$.

How does it work for cryo-EM?

- We scan the micrographs with a sliding window of size $L \times L$.
- We compute the first three autocorrelations of each window with respect to the center point and average over all windows.

How does it work for cryo-EM?

- We scan the micrographs with a sliding window of size $L \times L$.
- We compute the first three autocorrelations of each window with respect to the center point and average over all windows.
- The autocorrelations of the micrographs converge to a scaled, biased version of the volume's autocorrelations:

$$\lim_{N\to\infty} a_y^1 = \gamma \left\langle a^1_{P_\omega(x)} \right\rangle_\omega ,$$

$$\lim_{N\to\infty} a_y^2[\ell_1,\ell_2] = \gamma \left\langle a^2_{P_\omega(x)}[\ell_1,\ell_2] \right\rangle_\omega + \text{bias},$$

$$\lim_{N\to\infty} a_y^3[\ell_1,\ell_2;\ell_3,\ell_4] = \gamma \left\langle a^3_{P_\omega(x)}[\ell_1,\ell_2;\ell_3,\ell_4] \right\rangle_\omega + \text{bias}.$$

How does it work for cryo-EM?

- We scan the micrographs with a sliding window of size $L \times L$.
- We compute the first three autocorrelations of each window with respect to the center point and average over all windows.
- The autocorrelations of the micrographs converge to a scaled, biased version of the volume's autocorrelations:

$$\lim_{N\to\infty} a_y^1 = \gamma \left\langle a^1_{P_\omega(x)} \right\rangle_\omega ,$$
$$\lim_{N\to\infty} a_y^2[\ell_1,\ell_2] = \gamma \left\langle a^2_{P_\omega(x)}[\ell_1,\ell_2] \right\rangle_\omega + \text{bias},$$
$$\lim_{N\to\infty} a_y^3[\ell_1,\ell_2;\ell_3,\ell_4] = \gamma \left\langle a^3_{P_\omega(x)}[\ell_1,\ell_2;\ell_3,\ell_4] \right\rangle_\omega + \text{bias}.$$

- No detection is required!

Recovery of the volume from its autocorrelations

- The third-order autocorrelation contains $\sim L^3$ independent cubic equations (rather than L^4) that can be related to the $\sim L^3$ coefficients of the volume.

Recovery of the volume from its autocorrelations

- The third-order autocorrelation contains $\sim L^3$ independent cubic equations (rather than L^4) that can be related to the $\sim L^3$ coefficients of the volume.

- We estimate the volume's coefficients and γ by least-squares.

Recovery of the volume from its autocorrelations

- The third-order autocorrelation contains $\sim L^3$ independent cubic equations (rather than L^4) that can be related to the $\sim L^3$ coefficients of the volume.

- We estimate the volume's coefficients and γ by least-squares.

- In the absence of noise (empirically): unique mapping between the volume and γ, and the first three autocorrelations of the data.

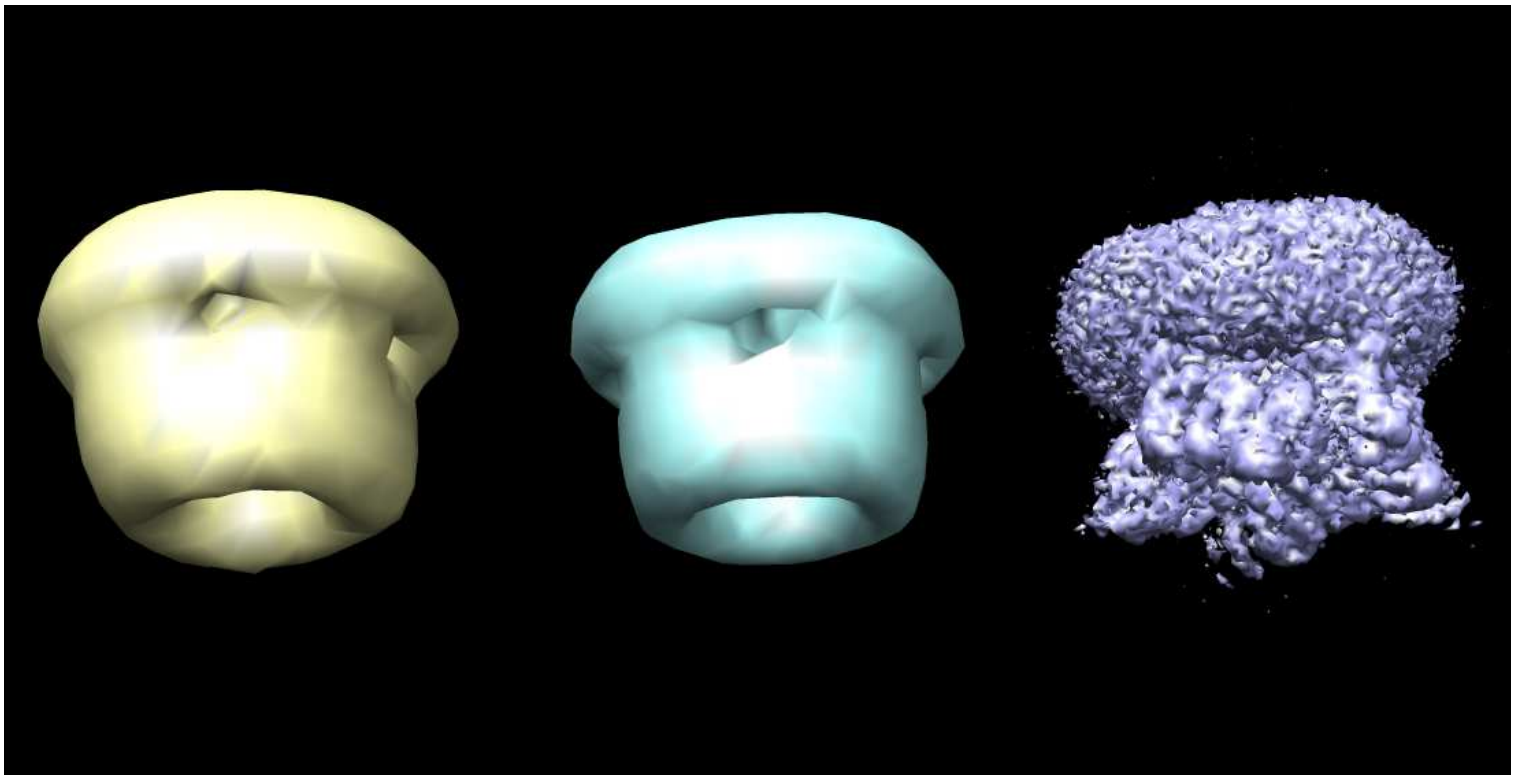
Recovery of the volume from its autocorrelations

- The third-order autocorrelation contains $\sim L^3$ independent cubic equations (rather than L^4) that can be related to the $\sim L^3$ coefficients of the volume.
- We estimate the volume's coefficients and γ by least-squares.
- In the absence of noise (empirically): unique mapping between the volume and γ, and the first three autocorrelations of the data.
- Unfortunately, the mapping is highly ill-conditioned, preventing stable recovery from noisy data.

Recovery of the volume from its autocorrelations

- The third-order autocorrelation contains $\sim L^3$ independent cubic equations (rather than L^4) that can be related to the $\sim L^3$ coefficients of the volume.
- We estimate the volume's coefficients and γ by least-squares.
- In the absence of noise (empirically): unique mapping between the volume and γ, and the first three autocorrelations of the data.
- Unfortunately, the mapping is highly ill-conditioned, preventing stable recovery from noisy data.
- Solution: Fourth-order autocorrelation! (Future work)

Recovery from clean autocorrelations



estimated structure (yellow), low-resolution structure (blue), high-resolution structure (purple)

TRPV1, the low-resolution molecule ($L = 5$) was down-sampled from 192^3 to 20^3 pixels

Outline

Future work

Cryo-EM:

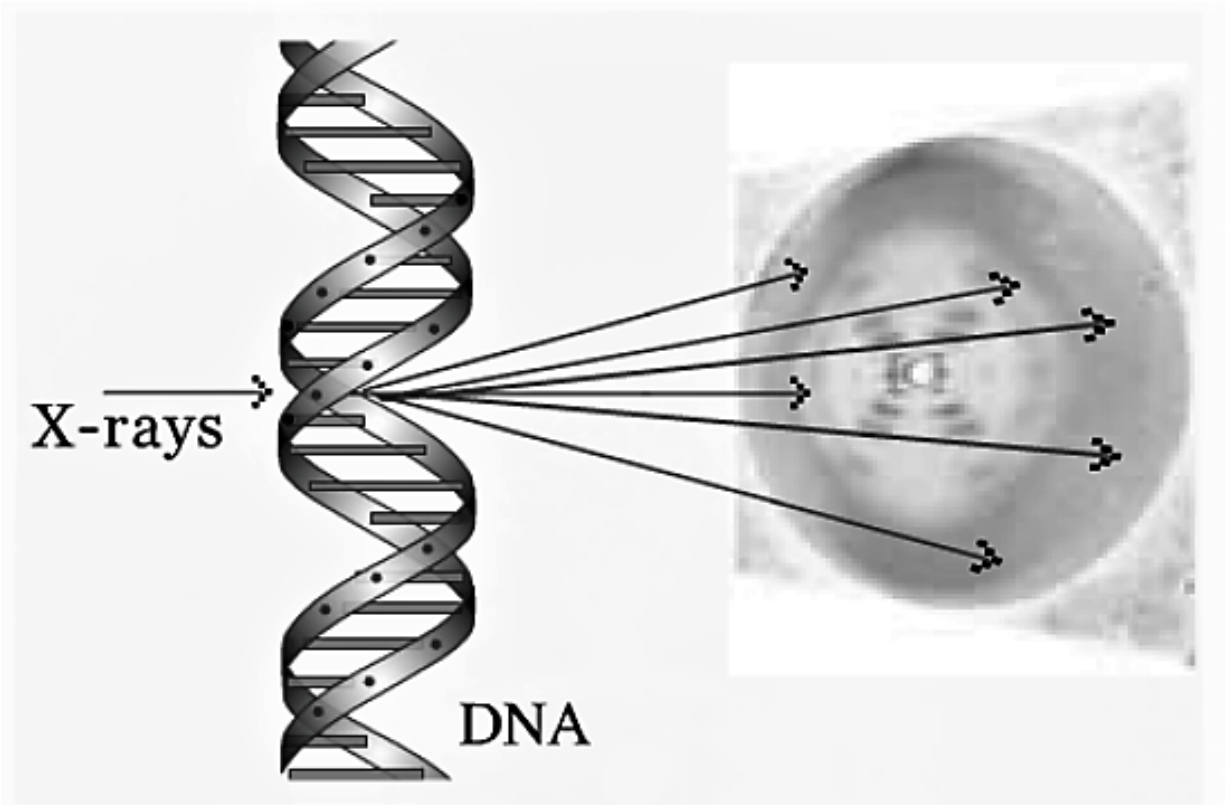
- Devising a full computational pipeline that produces high resolution 3-D structures directly from the micrograph:
  - Extending the framework to the fourth-order autocorrelation
  - A more accurate model
- 2-D classification (Ma, Bendory, Boumal, Sigworth, Singer, '18)
- What is the sample complexity of cryo-EM?

Signal processing/optimization/statistics:

- Efficient moment estimation
- The success of non-convex programs
- Heterogeneous models of MRA and blind deconvolution

Phase retrieval

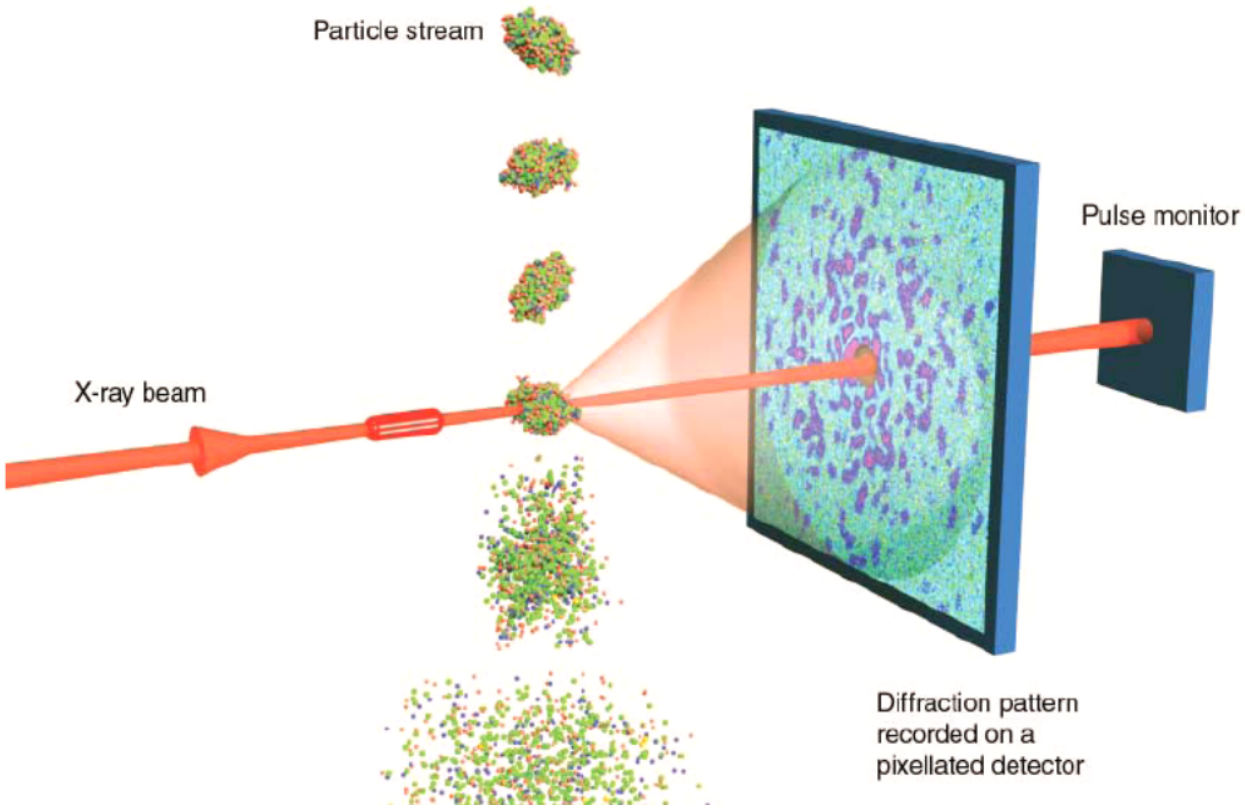
Phase retrieval is the problem of recovering a signal from its Fourier magnitudes.



Uncovering the double helix structure of the DNA with X-ray crystallography in 1951. Nobel Prize for Watson, Crick, and Wilkins in 1962 based on work by Rosalind Franklin.

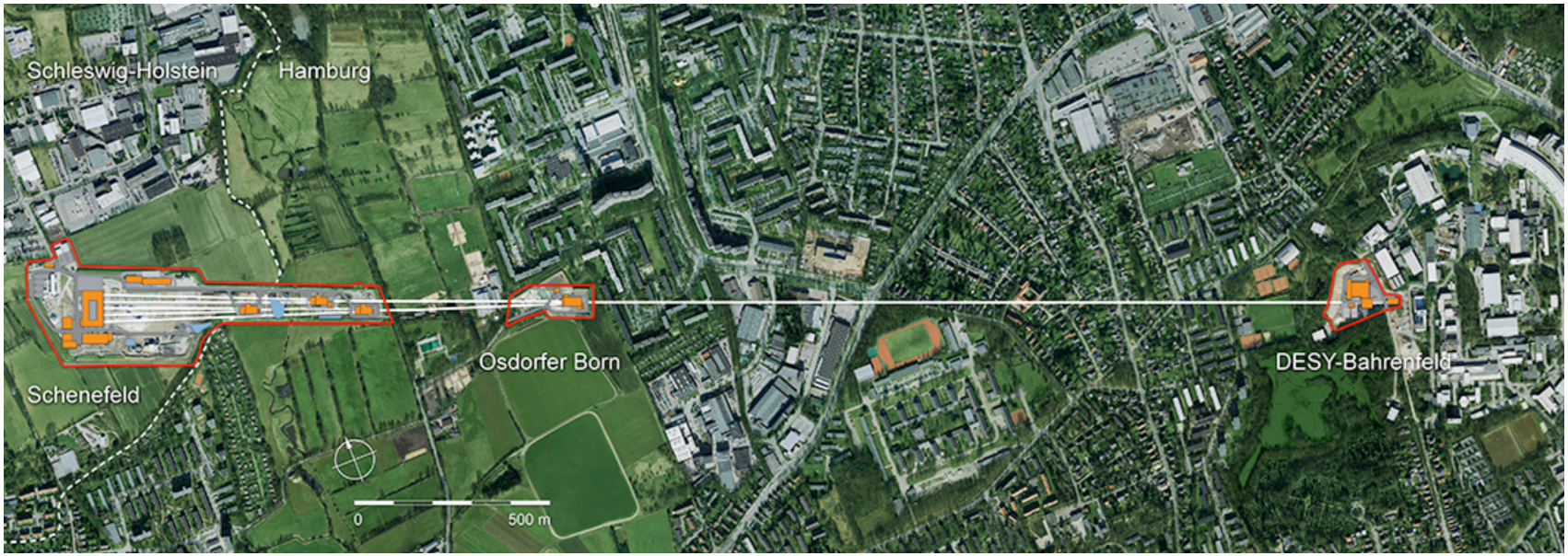
Single particle reconstruction using X-ray free-electron laser (XFEL)

XFEL $\approx$ cryo-EM + phase retrieval



Picture credit: (Gaffney and Chapman, '07)

European XFEL



- 3.4 kilometre-long facility
- User operation began in September 2017.
- 12 countries are participating in the project: Denmark, France, Germany, Hungary, Italy, Poland, Russia, Slovakia, Spain, Sweden, Switzerland, and the United Kingdom.
- The construction costs amount to 1.25 billion euro.

References

- Bendory, Boumal, Leeb, Levin, and Singer. "Toward single particle reconstruction without particle picking: Breaking the detection limit." Available at *arXiv preprint arXiv:1810.00226*, 2018.

- Ma, Bendory, Boumal, Sigworth, and Singer. "Heterogeneous multireference alignment for images with application to 2-D classification in single particle reconstruction." Available at *arXiv preprint arXiv:1811.10382*, 2018.

- Levin, Bendory, Boumal, Kileel, and Singer. "3D ab initio modeling in cryo-EM by autocorrelation analysis." *ISBI*, 2018.

- Boumal, Bendory, Lederman, and Singer. "Heterogeneous multireference alignment: A single pass approach." *CISS*, 2018.

- Abbe, Bendory, Leeb, Pereira, Sharon, and Singer. "Multireference alignment is easier with an aperiodic translation distribution." To appear in *IEEE Transactions on Information Theory*, 2018.

- Bendory, Boumal, Ma, Zhao, and Singer. "Bispectrum inversion with application to multireference alignment." *IEEE Transactions on Signal Processing* 66, no. 4 (2017): 1037-1050.

Thanks for your attention!