Estimation below the identifiability limit with application to cryo–EM

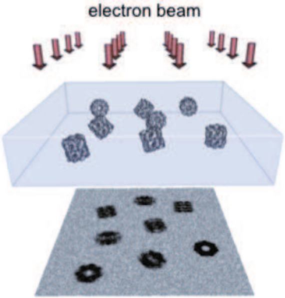
Tamir Bendory

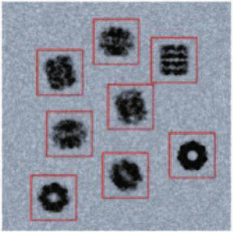
January 23, 2018

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

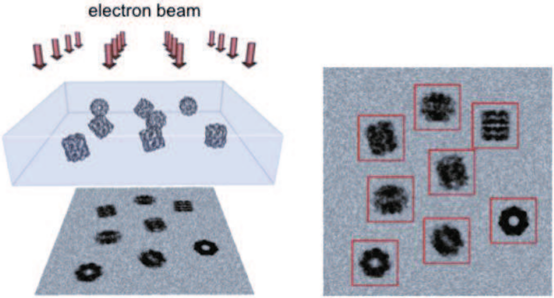
Table of contents

Single particle reconstruction using cryo–EM





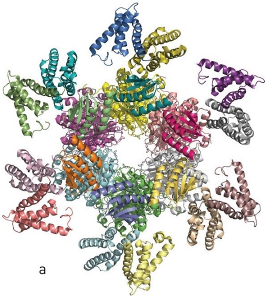
Single particle reconstruction using cryo–EM

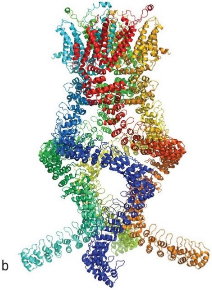


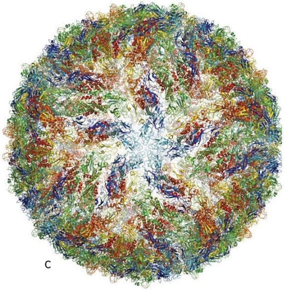
Why cryo–EM?

- Mapping the structure of molecules without crystallizing them
- Imaging of heterogeneous samples, with mixtures of molecules or multiple conformations

The resolution revolution







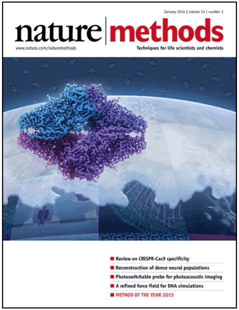
A protein complex that governs the circadian rhythm (sleep/wake cycle)

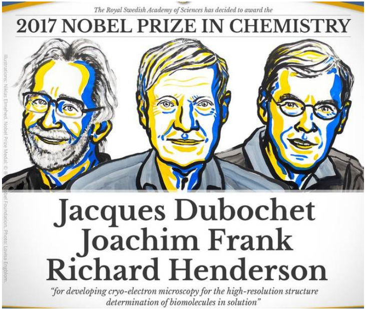
A sensor of the type that reads pressure changes in the ear

The Zika virus

https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2017/

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"

Image formation model and inverse problem

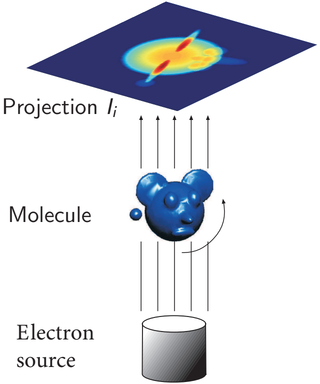


Image formation model

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

Image formation model and inverse problem

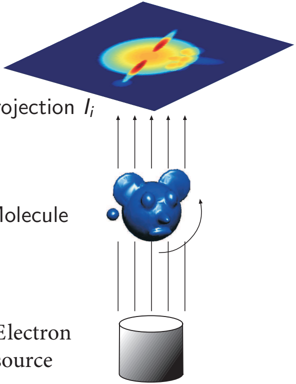


Image formation model

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

The cryo–EM problem

Estimate X given $I_1, \ldots, I_N$

Image formation model and inverse problem

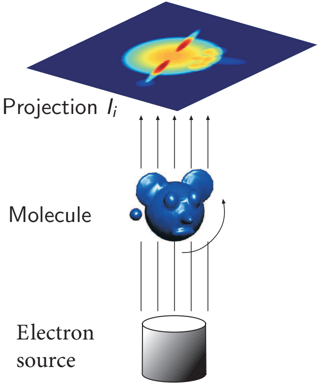


Image formation model

$I_i = P(R_i \circ X) + \mathrm{noise}, \ R_i \in SO(3)$

The cryo–EM problem

Estimate X given $I_1, \ldots, I_N$

The heterogeneity problem

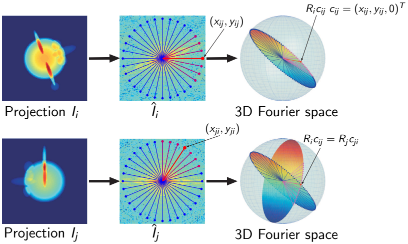
Estimate $X_1, \ldots, X_K$ given $I_1, \ldots, I_N$

Fundamental challenges

1. Unknown viewing directions

Fundamental challenges

1. Unknown viewing directions

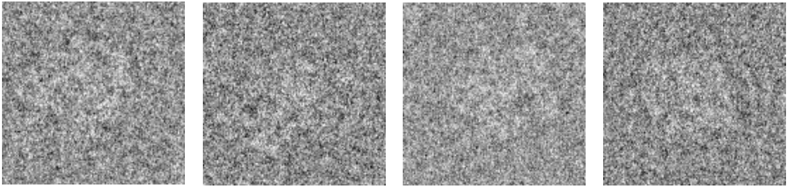


[Singer and Shkolnisky (2011)]

Fundamental challenges

1. Unknown viewing directions

2. Challenging SNR regime



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

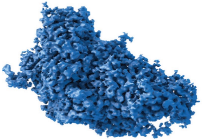
Fundamental challenges

1. Unknown viewing directions

2. Challenging SNR regime

3. Massive datasets

2.2 Å resolution $\Longleftrightarrow$ 12.4 TB



Bartesaghi et al. "2.2 Åresolution cryo-EM structure of β–galactosidase in complex with a cell-permeant inhibitor."

Table of contents

The multireference alignment problem

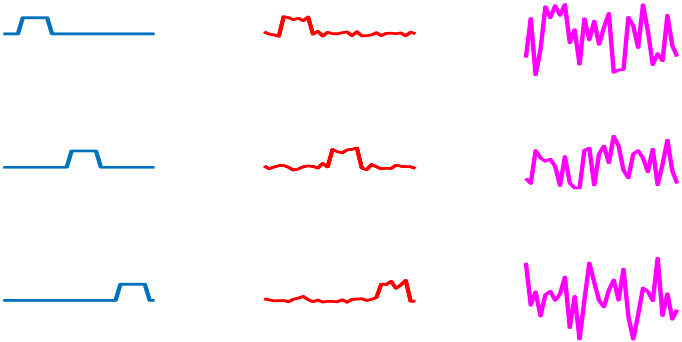
Problem: Estimating a signal $x \in \mathbb{R}^L$, up to cyclic–translation, from its noisy circularly–translated copies

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

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

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

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



Connection with the cryo–EM problem

The problem is mainly inspired by the cryo–EM problem:

$$y_i = P(R_i \circ X) + \text{noise},\ X \in \mathbb{R}^3,\ R_i \in SO(3)$$

Connection with the cryo–EM problem

The problem is mainly inspired by the cryo–EM problem:

$$y_i = P(R_i \circ X) + \text{noise},\ X \in \mathbb{R}^3,\ R_i \in SO(3)$$

	Cryo–EM	**MRA**
Signal	3D continuous signal	1D discrete signal
Latent variables	3D rotations	cyclic translations
Linear operator	tomographic projection	no projection
Interesting regime	low SNR, large N	low SNR, large N

Connection with the cryo–EM problem

The problem is mainly inspired by the cryo–EM problem:

$$y_i = P(R_i \circ X) + \text{noise}, \; X \in \mathbb{R}^3, \; R_i \in SO(3)$$

	Cryo–EM	**MRA**
Signal	3D continuous signal	1D discrete signal
Latent variables	3D rotations	cyclic translations
Linear operator	tomographic projection	no projection
Interesting regime	low SNR, large N	low SNR, large N

Connection with the cryo–EM problem

The problem is mainly inspired by the cryo–EM problem:

$$y_i = P(R_i \circ X) + \text{noise}, \; X \in \mathbb{R}^3, \; R_i \in SO(3)$$

	Cryo–EM	**MRA**
Signal	3D continuous signal	1D discrete signal
Latent variables	3D rotations	cyclic translations
Linear operator	tomographic projection	no projection
Interesting regime	low SNR, large N	low SNR, large N

Connection with the cryo–EM problem

The problem is mainly inspired by the cryo–EM problem:

$$y_i = P(R_i \circ X) + \text{noise},\ X \in \mathbb{R}^3,\ R_i \in SO(3)$$

	Cryo–EM	**MRA**
Signal	3D continuous signal	1D discrete signal
Latent variables	3D rotations	cyclic translations
Linear operator	tomographic projection	no projection
Interesting regime	low SNR, large N	low SNR, large N

Multireference alignment via alignment

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 translations, we can estimate

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

This is an unbiased estimator with variance σ^2/N.

Multireference alignment via alignment

Model

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

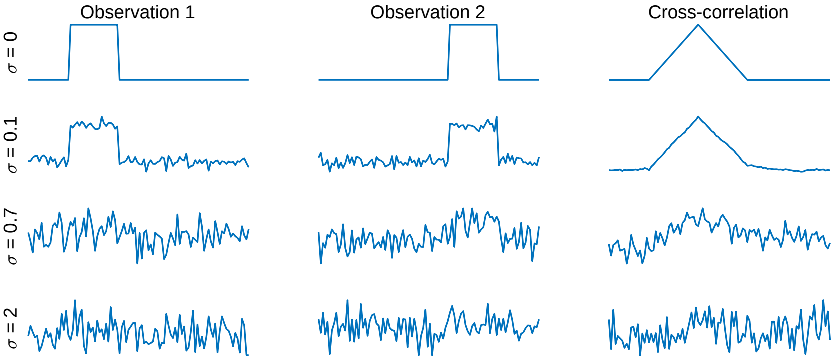
Given the translations, we can estimate

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

This is an unbiased estimator with variance σ^2/N.

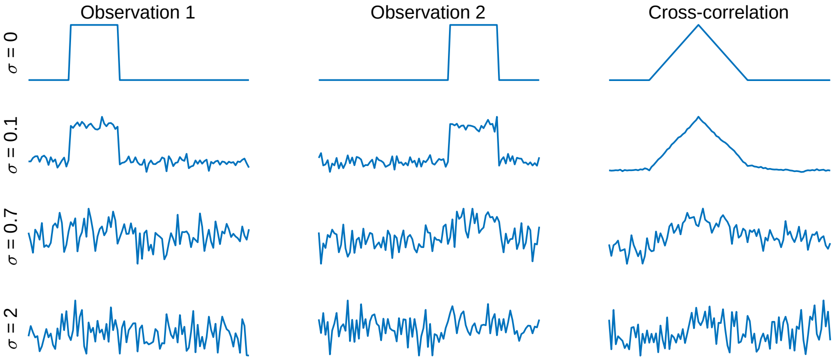
Can we estimate the translations?

Alignment



Observation 1
Observation 2
Cross-correlation
σ = 0
σ = 0.1
σ = 0.7
σ = 2

Alignment



Alignment is impossible in the low SNR regime!

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0, \ldots, L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0, \ldots, L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]
  - For almost any other translation distribution, $N \gtrsim \sigma^4$. [Abbe, B, Leeb, Pereira, Sharon, Singer (2017)]

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0, \ldots, L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]
  - For almost any other translation distribution, $N \gtrsim \sigma^4$. [Abbe, B, Leeb, Pereira, Sharon, Singer (2017)]
- It is possible to accurately reconstruct the signal from sufficiently many noisy shifted copies for arbitrarily low SNR.

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0,\ldots,L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]
  - For almost any other translation distribution, $N \gtrsim \sigma^4$. [Abbe, B, Leeb, Pereira, Sharon, Singer (2017)]
- It is possible to accurately reconstruct the signal from sufficiently many noisy shifted copies for arbitrarily low SNR.
- Note that if the shifts are known, then $N \gtrsim \sigma^2$. The fact that shifts are not known has a big impact.

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0, \ldots, L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]
  - For almost any other translation distribution, $N \gtrsim \sigma^4$. [Abbe, B, Leeb, Pereira, Sharon, Singer (2017)]
- It is possible to accurately reconstruct the signal from sufficiently many noisy shifted copies for arbitrarily low SNR.
- Note that if the shifts are known, then $N \gtrsim \sigma^2$. The fact that shifts are not known has a big impact.

Information-theoretic limits

Can we reconstruct the signal accurately while estimating most shifts poorly?

- Lower bounds:
  - If $r_i \sim U[0, \ldots, L-1]$, then $N \gtrsim \sigma^6$. [Bandeira, Rigollet, Weed (2017)]
  - For almost any other translation distribution, $N \gtrsim \sigma^4$. [Abbe, B, Leeb, Pereira, Sharon, Singer (2017)]
- It is possible to accurately reconstruct the signal from sufficiently many noisy shifted copies for arbitrarily low SNR.
- Note that if the shifts are known, then $N \gtrsim \sigma^2$. The fact that shifts are not known has a big impact.

Can we achieve the optimal estimation rate?

Expectation–maximization

An iterative method to find maximum marginal likelihood:

$$x_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_k^{\ell,i}R_\ell^{-1}y_i, \quad \omega_k^{\ell,i} \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_k - y_i\|_2^2\right)$$

Expectation–maximization

An iterative method to find maximum marginal likelihood:

$$x_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_k^{\ell,i}R_\ell^{-1}y_i, \quad \omega_k^{\ell,i} \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_k - y_i\|_2^2\right)$$

Good numerical performance

Expectation–maximization

An iterative method to find maximum marginal likelihood:

$$x_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_k^{\ell,i}R_\ell^{-1}y_i, \quad \omega_k^{\ell,i} \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_k - y_i\|_2^2\right)$$

Good numerical performance

At poor SNR, requires many iterations

Expectation–maximization

An iterative method to find maximum marginal likelihood:

$$x_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L-1}\omega_k^{\ell,i}R_\ell^{-1}y_i, \quad \omega_k^{\ell,i} \propto exp\left(-\frac{1}{2\sigma^2}\|R_\ell x_k - y_i\|_2^2\right)$$

Good numerical performance

At poor SNR, requires many iterations

Can we achieve similar performance with only **one pass** over the data?

Cyclic-shift invariant features

We need the first three moments to determine a signal

$$\text{(mean)} \quad \mu_x := \hat{x}[0]/L$$

$$\text{(power spectrum)} \quad P_x[k] := \hat{x}[k]\overline{\hat{x}[k]} = |\hat{x}[k]|^2$$

$$\text{(bispectrum)} \quad B_x[k,\ell] := \hat{x}[k]\overline{\hat{x}[\ell]}\hat{x}[\ell - k]$$

Algorithms and analysis are provided in "Bispectrum inversion with application to multireference alignment", Bendory, Boumal, Ma, Zhao and Singer.

Cyclic-shift invariant features

We need the first three moments to determine a signal

$$\text{(mean)} \quad \mu_x := \hat{x}[0]/L$$

$$\text{(power spectrum)} \quad P_x[k] := \hat{x}[k]\overline{\hat{x}[k]} = |\hat{x}[k]|^2$$

$$\text{(bispectrum)} \quad B_x[k,\ell] := \hat{x}[k]\overline{\hat{x}[\ell]}\hat{x}[\ell-k]$$

Stable recovery using a non-convex least-squares

$$\min_z |\mu_z - \tilde{\mu}_x|^2 + \lambda_1 \left\| P_z - \tilde{P}_x \right\|_2^2 + \lambda_2 \left\| B_z - \tilde{B}_x \right\|_F^2$$

Algorithms and analysis are provided in "Bispectrum inversion with application to multireference alignment", Bendory, Boumal, Ma, Zhao and Singer.

Multireference alignment by invariant features

The invariant features can be estimated from the data:

$$M_1 := \frac{1}{N}\sum_{i=1}^{N} \mu_{y_i} \to \mu_x, \quad \mathrm{Var}\,(M_1) \sim \sigma^2/N,$$

$$M_2 := \frac{1}{N}\sum_{i=1}^{N} P_{y_i} \to P_x + \sigma^2 L\mathbf{1}, \quad \mathrm{Var}\,(M_2) \sim \sigma^4/N,$$

$$M_3 := \frac{1}{N}\sum_{i=1}^{N} B_{y_i} \to B_x + \mu_x\sigma^2 L^2 A, \quad \mathrm{Var}\,(M_3) \sim \sigma^6/N.$$

Multireference alignment by invariant features

The invariant features can be estimated from the data:

$$M_1 := \frac{1}{N}\sum_{i=1}^{N} \mu_{y_i} \to \mu_x, \quad \mathrm{Var}\left(M_1\right) \sim \sigma^2/N,$$

$$M_2 := \frac{1}{N}\sum_{i=1}^{N} P_{y_i} \to P_x + \sigma^2 L\mathbf{1}, \quad \mathrm{Var}\left(M_2\right) \sim \sigma^4/N,$$

$$M_3 := \frac{1}{N}\sum_{i=1}^{N} B_{y_i} \to B_x + \mu_x\sigma^2 L^2 A, \quad \mathrm{Var}\left(M_3\right) \sim \sigma^6/N.$$

Achieving the optimal estimation rate σ^6/N

Multireference alignment by invariant features

The invariant features can be estimated from the data:

$$M_1 := \frac{1}{N}\sum_{i=1}^{N}\mu_{y_i} \to \mu_x, \quad \mathrm{Var}\left(M_1\right) \sim \sigma^2/N,$$

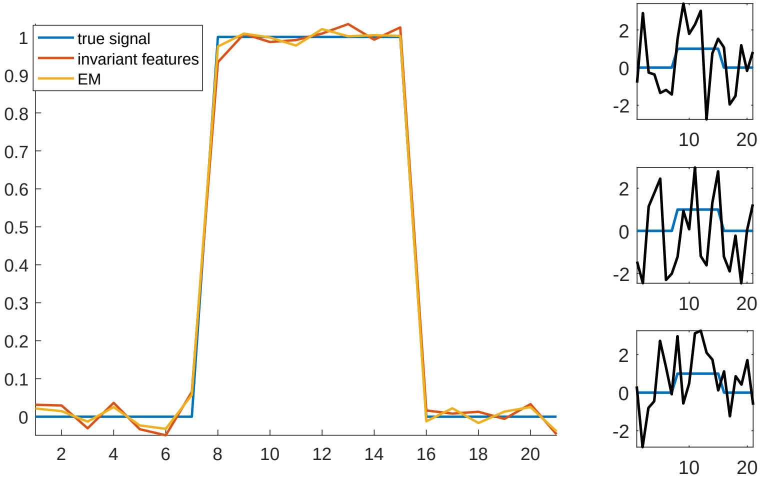
$$M_2 := \frac{1}{N}\sum_{i=1}^{N}P_{y_i} \to P_x + \sigma^2 L\mathbf{1}, \quad \mathrm{Var}\left(M_2\right) \sim \sigma^4/N,$$

$$M_3 := \frac{1}{N}\sum_{i=1}^{N}B_{y_i} \to B_x + \mu_x\sigma^2 L^2 A, \quad \mathrm{Var}\left(M_3\right) \sim \sigma^6/N.$$

Achieving the optimal estimation rate σ^6/N

Computational complexity linear in N (requires only one pass over the data)

Numerical example



$N = 10^5$, $\sigma = 1.5$. Running time: invariants features = 2.1 [sec], EM = 67.2 [sec]

Multireference alignment with heterogeneity

Problem: Estimating a set of signals $x_1, \ldots, x_K$ from their noisy, unlabeled, circularly-translated copies

$$y_i = R_{r_i} x_{v_i} + \varepsilon_i, \quad i = 1, \ldots, N.$$

$v_i = k$ with probability w_i, $w \in \Delta^K$.

Boumal, Bendory, Lederman and Singer. "Heterogeneous multireference alignment: a single pass approach."

Multireference alignment with heterogeneity

Problem: Estimating a set of signals $x_1, \ldots, x_K$ from their noisy, unlabeled, circularly-translated copies

$$y_i = R_{r_i} x_{v_i} + \varepsilon_i, \quad i = 1, \ldots, N.$$

$v_i = k$ with probability w_i, $w \in \Delta^K$.

In the low SNR regime, clustering is also impossible

Boumal, Bendory, Lederman and Singer. "Heterogeneous multireference alignment: a single pass approach."

Multireference alignment with heterogeneity

Problem: Estimating a set of signals $x_1, \ldots, x_K$ from their noisy, unlabeled, circularly-translated copies

$$y_i = R_{r_i} x_{v_i} + \varepsilon_i, \quad i = 1, \ldots, N.$$

$v_i = k$ with probability w_i, $w \in \Delta^K$.

In the low SNR regime, clustering is also impossible

Can we reconstruct the signals accurately while estimating most shifts and clusters poorly?

Boumal, Bendory, Lederman and Singer. "Heterogeneous multireference alignment: a single pass approach."

Mixed invariant feature

We can estimate the **mixed** invariant features from the data:

$$M_1 := \frac{1}{N}\sum_{j=1}^{N}\mu_{y_j} \to \sum_{i=1}^{k} w_i \mu_{x_i},$$

$$M_2 := \frac{1}{N}\sum_{j=1}^{N}P_{y_j} \to \sum_{i=1}^{k} w_i P_{x_i} + \sigma^2 L\mathbf{1},,$$

$$M_3 := \frac{1}{N}\sum_{j=1}^{N}B_{y_j} \to \sum_{i=1}^{k} w_i B_{x_i} + \left(\sum_{i=1}^{k} w_i \mu_{x_i}\right)\sigma^2 L^2 A.$$

Mixed invariant feature

We can estimate the **mixed** invariant features from the data:

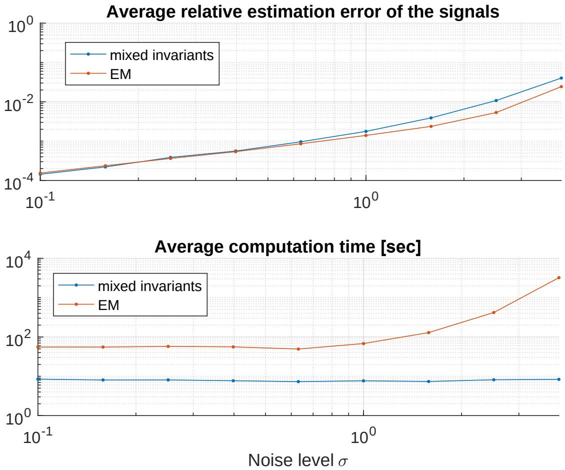
$$M_1 := \frac{1}{N}\sum_{j=1}^{N}\mu_{y_j} \to \sum_{i=1}^{k} w_i \mu_{x_i},$$

$$M_2 := \frac{1}{N}\sum_{j=1}^{N}P_{y_j} \to \sum_{i=1}^{k} w_i P_{x_i} + \sigma^2 L\mathbf{1},,$$

$$M_3 := \frac{1}{N}\sum_{j=1}^{N}B_{y_j} \to \sum_{i=1}^{k} w_i B_{x_i} + \left(\sum_{i=1}^{k} w_i \mu_{x_i}\right)\sigma^2 L^2 A.$$

Signal estimation by non-convex least-squares

Numerical results



$L = 50$, $K = 2$ with i.i.d. normal entries, $N = 2 \times 10^6$, 20 repetitions

Table of contents

2D classification for cryo–EM

Goal: Given N noisy cryo–EM images, estimate accurately $K \ll N$ representative images

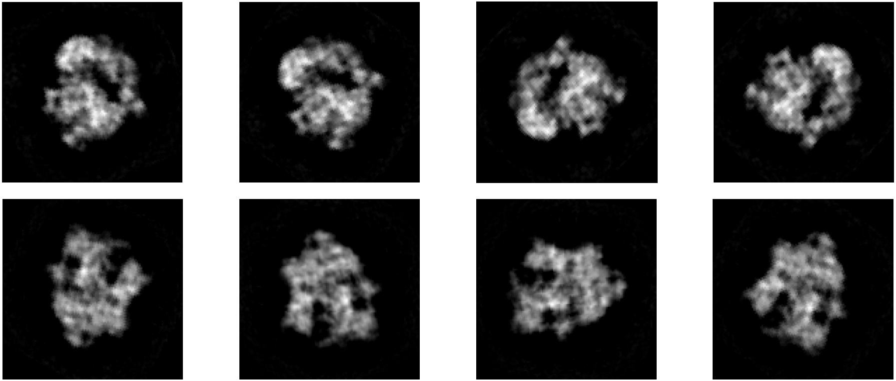
---

Ongoing project with Chao Ma, Nicolas Boumal and Amit Singer

2D classification for cryo–EM

Goal: Given N noisy cryo–EM images, estimate accurately $K \ll N$ representative images

Simplified example: Set of rotated copies of 2 images

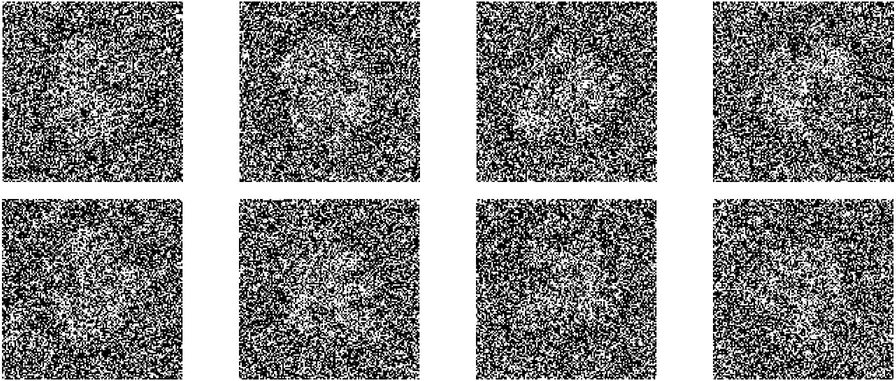


Ongoing project with Chao Ma, Nicolas Boumal and Amit Singer

2D classification for cryo–EM

Simplified example: Estimate 2 images, up to rotation, from their noisy rotated copies

Noisy rotated images: (SNR $= 1/50$):



Invariants of a steerable basis

Each image can be expanded by a **steerable** basis

$$X(r,\theta) = \sum_{k,q} A_{k,q} u^{k,q}(r,\theta),$$

satisfying

$$X(r,\theta-\alpha) = \sum_{k,q} A_{k,q} e^{-\iota k\alpha} u^{k,q}(r,\theta).$$

Invariants of a steerable basis

Each image can be expanded by a **steerable** basis

$$X(r,\theta) = \sum_{k,q} A_{k,q} u^{k,q}(r,\theta),$$

satisfying

$$X(r,\theta-\alpha) = \sum_{k,q} A_{k,q} e^{-\iota k\alpha} u^{k,q}(r,\theta).$$

The **rotationally-invariant** features are:

$$\begin{aligned} \text{(mean)} \quad & A_{0,q}, \quad \forall q \\ \text{(power spectrum)} \quad & A_{k,q_1}\overline{A_{k,q_2}}, \quad \forall k, q_1, q_2 \\ \text{(bispectrum)} \quad & A_{k_1,q_1}\overline{A_{k_2,q_2}}A_{k_2-k_1,q_3}, \quad \forall k_1, k_2, q_1, q_2, q_3 \end{aligned}$$

Algorithm for 2D classification

1. Expand each image in a steerable basis
2. Compute the rotationally-invariant features
3. Average over all images to estimate the **mixed** invariant features
4. Solve a non-convex least-squares

Numerical example

Estimated images (5000 images per class, SNR $= 1/50$):

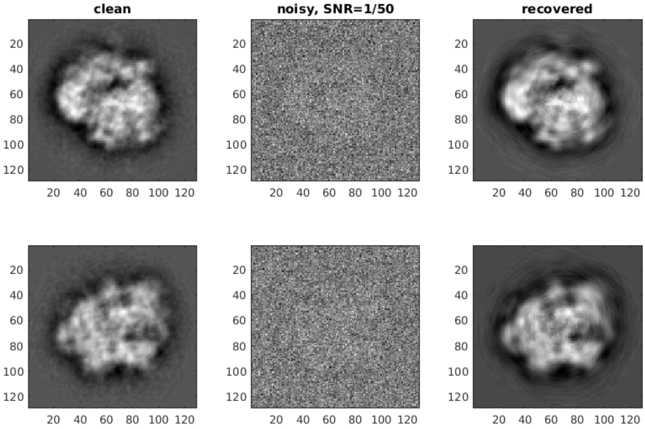


Table of contents

The invariants of the cryo–EM problem

Theorem (Levin, B, Boumal, Kileel, Singer)

If the viewing directions are drawn from the uniform distribution, then the second moment of the projections and two clean images determine the molecule, up to global rotation.

- Requires one pass over the data (very fast, computational complexity linear in N)

The invariants of the cryo–EM problem

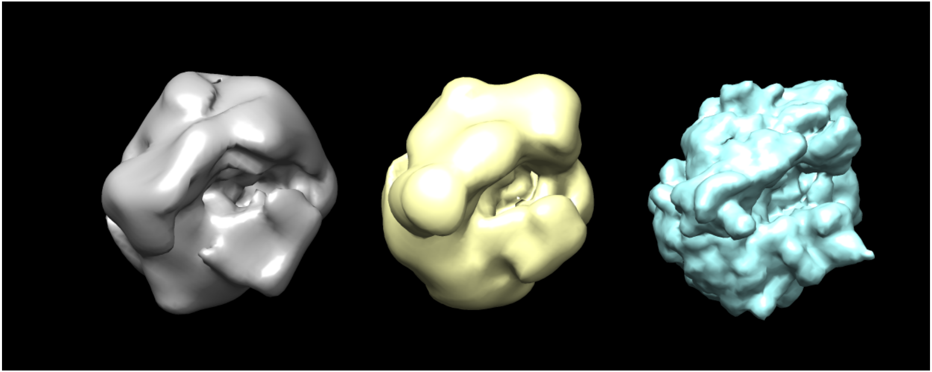
Theorem (Levin, B, Boumal, Kileel, Singer)

If the viewing directions are drawn from the uniform distribution, then the second moment of the projections and two clean images determine the molecule, up to global rotation.

- Requires one pass over the data (very fast, computational complexity linear in N)
- These conditions are not met in practice and therefore only low-resolution estimation is possible

3D ab-initio modeling

EMDB 5360, $50,000$ projections, $\text{SNR} = 1/10$

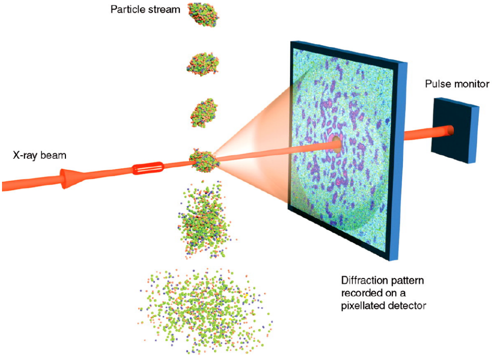


estimation | low-resolution "ground truth" | "ground truth"

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

Single Particle Reconstruction using X-ray Free-Electron Laser (XFEL)

XFEL $\approx$ cryo–EM + phase retrieval



[Gaffney and Chapman (2007)]

Main theoretical questions

Non-convex optimization

Why is it so effective for cryo–EM and multireference alignment?

Reasons for optimism: Recent developments in statistical estimation problems, such as phase retrieval, blind deconvolution, matrix completion and phase synchronization

Main theoretical questions

Non-convex optimization

Why is it so effective for cryo–EM and multireference alignment?

Reasons for optimism: Recent developments in statistical estimation problems, such as phase retrieval, blind deconvolution, matrix completion and phase synchronization

Information-theoretic limit

What is the sample complexity of the cryo–EM problem?

Partial results:

- Levin, Bendory, Boumal, Kileel and Singer. "3D ab initio modeling in cryo-EM by autocorrelation analysis."
- Bandeira et al. "Estimation under group actions: recovering orbits from invariants." *arXiv:1712.10163* (2017).

References

- Tamir Bendory , Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. "Bispectrum Inversion with Application to Multireference Alignment". *IEEE Transactions on Signal Processing*, vol. 66, issue 4, pp. 1037-1050, 2018.

- Nicolas Boumal, Tamir Bendory, Roy R Lederman and Amit Singer. "Heterogeneous multireference alignment: a single pass approach". *arXiv preprint arXiv:1710.02590* (2017).

- Emmanuel Abbe, Tamir Bendory, William Leeb, João Pereira Nir Sharon and Amit Singer. "Multireference Alignment is Easier with an Aperiodic Translation Distribution". *arXiv preprint arXiv:1710.02793* (2017).

- Eitan Levin, Tamir Bendory, Nicolas Boumal, Joseph Kileel and Amit Singer. "3-D ab-initio modeling in cryo-EM by autocorrelation analysis". *To appear in ISBI 2018*.

Thanks for your attention!