Some Optimization and Statistical Learning Problems in Structural Biology

Amit Singer

Princeton University, Department of Mathematics and PACM

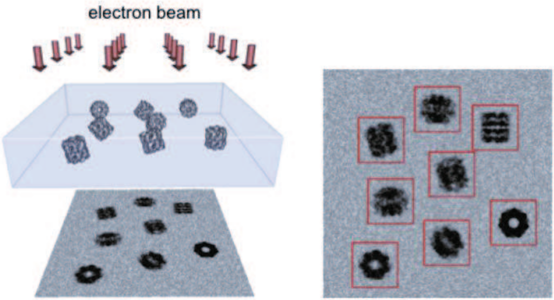
January 8, 2013

Outline / Advertisement

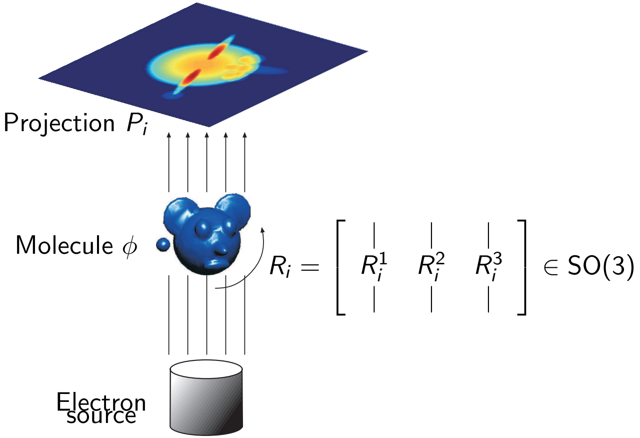
- Two alternative techniques to X-ray crystallography:
    1. Single particle cryo-electron microscopy
    2. Nuclear Magnetic Resonance (NMR) Spectroscopy
- Methods (a few examples of what is done now)
- Challenges
- Looking forward to your input
- Also looking for students and postdocs

Single Particle Cryo-Electron Microscopy

Drawing of the imaging process:

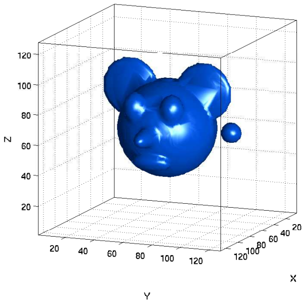


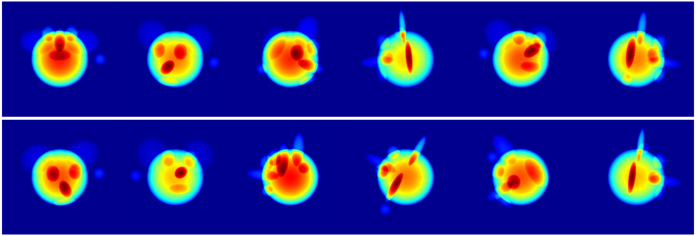
Single Particle Cryo-Electron Microscopy: Model



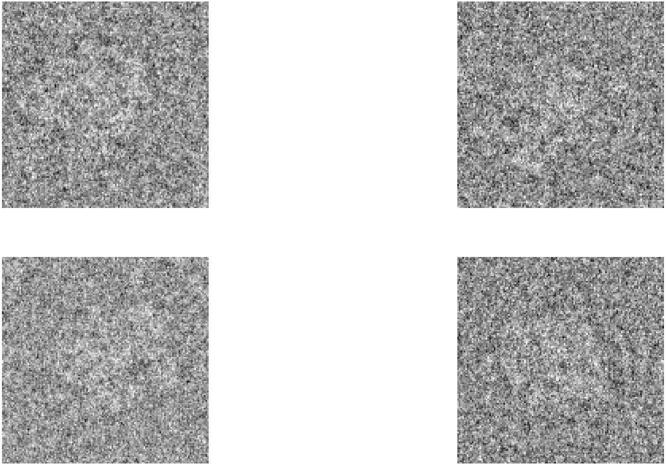
- Projection images $P_i(x, y) = \int_{-\infty}^{\infty} \phi(xR_i^1 + yR_i^2 + zR_i^3)\, dz +$ "noise".
- $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is the electric potential of the molecule.
- Cryo-EM problem: Find ϕ and $R_1, \ldots, R_n$ given $P_1, \ldots, P_n$.

Toy Example

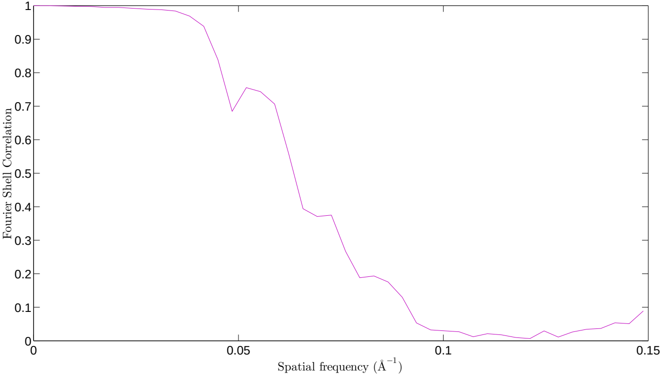




E. coli 50S ribosomal subunit: sample images
Fred Sigworth, Yale Medical School



Movie by Lanhui Wang and Zhizhen (Jane) Zhao



Algorithmic Pipeline

- **Particle Picking:** manual, automatic or experimental image segmentation.
- **Class Averaging:** classify images with similar viewing directions, register and average to improve their signal-to-noise ratio (SNR). S, Zhao, Shkolnisky, Hadani, SIIMS, 2011.
- **Orientation Estimation:** S, Shkolnisky, SIIMS, 2011.
- **Three-dimensional Reconstruction:** a 3D volume is generated by a tomographic inversion algorithm.
- **Iterative Refinement**

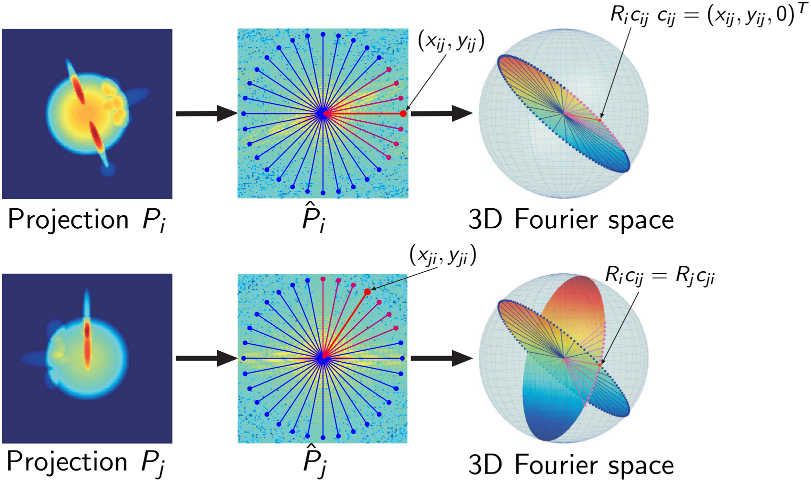
Assumptions for today's talk:

- Trivial point-group symmetry
- Homogeneity

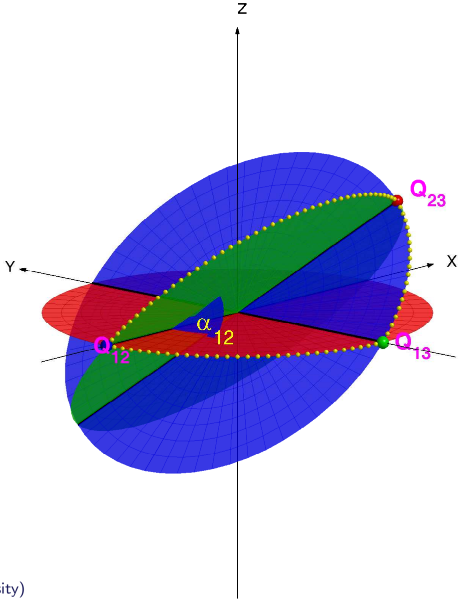
What mathematics do we use to solve the problem?

- Tomography
- Convex optimization and semidefinite programming
- Random matrix theory (in several places)
- Representation theory of SO(3)
  (if viewing directions are uniformly distributed)
- Spectral graph theory, (vector) diffusion maps
- Fast randomized algorithms
- ...

Orientation Estimation: Fourier projection-slice theorem



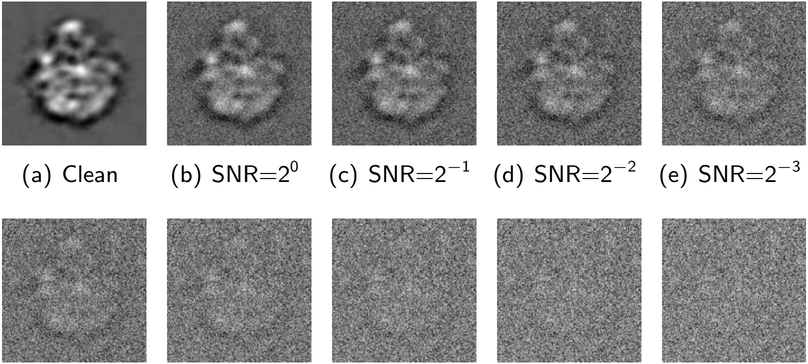
Angular Reconstitution (Van Heel 1987, Vainshtein and Goncharov 1986)



Experiments with simulated noisy projections

- Each projection is 129x129 pixels.

$$\text{SNR} = \frac{\text{Var}(Signal)}{\text{Var}(Noise)},$$



(a) Clean (b) SNR=2^{0} (c) SNR=2^{-1} (d) SNR=2^{-2} (e) SNR=2^{-3}

(f) SNR=2^{-4} (g) SNR=2^{-5} (h) SNR=2^{-6} (i) SNR=2^{-7} (j) SNR=2^{-8}

Fraction of correctly identified common lines and the SNR

- Define common line as being correctly identified if both radial lines deviate by no more than 10° from true directions.
- Fraction p of correctly identified common lines increases by PCA

$\log_2(\mathrm{SNR})$	p
20	0.997
0	0.980
-1	0.956
-2	0.890
-3	0.764
-4	0.575
-5	0.345
-6	0.157
-7	0.064
-8	0.028
-9	0.019

Least Squares Approach

- Consider the unit directional vectors as three-dimensional vectors:

$$
\begin{aligned}
c_{ij} &= (x_{ij}, y_{ij}, 0)^T, \\
c_{ji} &= (x_{ji}, y_{ji}, 0)^T.
\end{aligned}
$$

- Being the common-line of intersection, the mapping of c_{ij} by R_i must coincide with the mapping of c_{ji} by R_j: $(R_i, R_j \in SO(3))$

$$R_i c_{ij} = R_j c_{ji}, \text{ for } 1 \le i < j \le n.$$

- Least squares:

$$\min_{R_1, R_2, \ldots, R_n \in SO(3)} \sum_{i \neq j} \| R_i c_{ij} - R_j c_{ji} \|^2$$

- Non-convex... Exponentially large search space...

Quadratic Optimization Under Orthogonality Constraints

We approximate the solution to the least squares problem

$$\min_{R_1,R_2,\ldots,R_n\in SO(3)} \sum_{i\neq j} \|R_i c_{ij} - R_j c_{ji}\|^2$$

using SDP and rounding. Related to:

- Goemans-Williamson SDP relaxation for MAX-CUT
- Generalized Orthogonal Procrustes Problem (see, e.g., Nemirovski 2007)

"Robust" version – Least Unsquared Deviations:

$$\min_{R_1,R_2,\ldots,R_n\in SO(3)} \sum_{i\neq j} \|R_i c_{ij} - R_j c_{ji}\|$$

- Motivated by recent suggestions for "robust PCA"
- Also admits semidefinite relaxation
- Solved by alternating direction augmented Lagrangian method
- Less sensitive to misidentifications of common-lines (outliers)

Spectral Relaxation for Uniformly Distributed Rotations

$$\begin{bmatrix} | & | \\ R_i^1 & R_i^2 \\ | & | \end{bmatrix} = \begin{bmatrix} x_i^1 & x_i^2 \\ y_i^1 & y_i^2 \\ z_i^1 & z_i^2 \end{bmatrix}, \quad i = 1, \ldots, n.$$

- Define 3 vectors of length $2n$

$$\begin{aligned} x &= \begin{bmatrix} x_1^1 & x_1^2 & x_2^1 & x_2^2 & \cdots & x_n^1 & x_n^2 \end{bmatrix}^T \\ y &= \begin{bmatrix} y_1^1 & y_1^2 & y_2^1 & y_2^2 & \cdots & y_n^1 & y_n^2 \end{bmatrix}^T \\ z &= \begin{bmatrix} z_1^1 & z_1^2 & z_2^1 & z_2^2 & \cdots & z_n^1 & z_n^2 \end{bmatrix}^T \end{aligned}$$

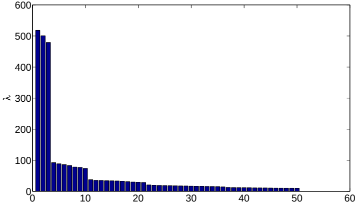
- Rewrite the least squares objective function as

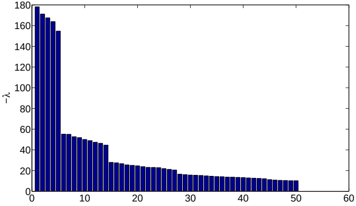
$$\max_{R_1,\ldots,R_n \in SO(3)} \sum_{i \neq j} \langle R_i c_{ij}, R_j c_{ji} \rangle = \max_{R_1,\ldots,R_n \in SO(3)} x^T C x + y^T C y + z^T C z$$

- By **symmetry**, if rotations are uniformly distributed over $SO(3)$, then the top eigenvalue of C has multiplicity 3 and corresponding eigenvectors are x, y, z from which we recover $R_1, R_2, \ldots, R_n$!

Spectrum of C

- Numerical simulation with $n = 1000$ rotations sampled from the Haar measure; no noise.
- Bar plot of positive (left) and negative (right) eigenvalues of C:





- Eigenvalues: $\lambda_l \approx n\frac{(-1)^{l+1}}{l(l+1)}, \quad l = 1, 2, 3, \ldots$ $(\frac{1}{2}, -\frac{1}{6}, \frac{1}{12}, \ldots)$
- Multiplicities: $2l + 1$.
- Two basic questions:
  1. Why this spectrum? Answer: Representation Theory of SO(3) (Hadani, S, 2011)
  2. Is it stable to noise? Answer: Yes, due to random matrix theory.

Probabilistic Model and Wigner's Semi-Circle Law

- **Simplistic Model:** every common line is detected correctly with probability p, independently of all other common-lines, and with probability $1-p$ the common lines are falsely detected and are uniformly distributed over the unit circle.
- Let C^{clean} be the matrix C when all common-lines are detected correctly $(p=1)$.
- The expected value of the noisy matrix C is
$$\mathbb{E}[C] = pC^{\text{clean}},$$
as the contribution of the falsely detected common lines to the expected value **vanishes**.
- Decompose C as
$$C = pC^{\text{clean}} + W,$$
where W is a $2n \times 2n$ zero-mean random matrix.
- The eigenvalues of W are distributed according to Wigner's semi-circle law whose support, up to $O(p)$ and finite sample fluctuations, is $[-\sqrt{2n}, \sqrt{2n}]$.

Threshold probability

- Sufficient condition for top three eigenvalues to be pushed away from the semi-circle and no other eigenvalue crossings: (rank-1 and finite rank deformed Wigner matrices, Füredi and Komlós 1981, Féral and Péché 2007, ...)

$$p\Delta(C^{\text{clean}}) > \frac{1}{2}\lambda_1(W)$$

- Spectral gap $\Delta(C^{\text{clean}})$ and spectral norm $\lambda_1(W)$ are given by

$$\Delta(C^{\text{clean}}) \approx (\frac{1}{2} - \frac{1}{12})n$$

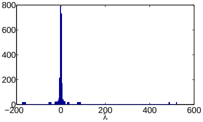
  and

$$\lambda_1(W) \approx \sqrt{2n}.$$

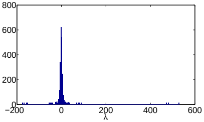
- Threshold probability

$$p_c = \frac{5\sqrt{2}}{6\sqrt{n}}.$$

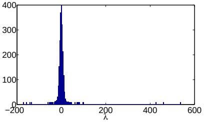
Numerical Spectra of C, $n = 1000$



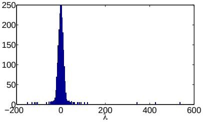
(a) SNR=2^0



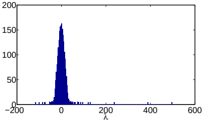
(b) SNR=2^{-1}



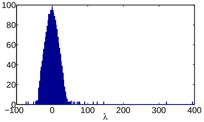
(c) SNR=2^{-2}



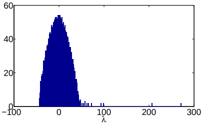
(d) SNR=2^{-3}



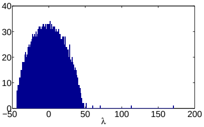
(e) SNR=2^{-4}



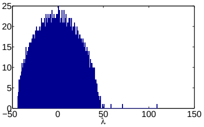
(f) SNR=2^{-5}



(g) SNR=2^{-6}



(h) SNR=2^{-7}



(i) SNR=2^{-8}

MSE for $n = 1000$

SNR	p	λ_1	λ_2	λ_3	λ_4	MSE
2^{-1}	0.951	523	491	475	89	0.0182
2^{-2}	0.890	528	490	450	92	0.0224
2^{-3}	0.761	533	482	397	101	0.0361
2^{-4}	0.564	530	453	307	119	0.0737
2^{-5}	0.342	499	381	193	134	0.2169
2^{-6}	0.168	423	264	133	101	1.8011
2^{-7}	0.072	309	155	105	80	2.5244
2^{-8}	0.032	210	92	86	70	3.5196

- Model fails at low SNR. Why?
- Wigner model is too simplistic – cannot have n^2 independent random variables from just n images.
- $C_{ij} = K(P_i, P_j)$, "kernel random matrix", related to Koltchinskii and Giné (2000), El-Karoui (2010)
- Kernel is discontinuous

Challenges / Work in Progress

- Currently not taking into account all available information: e.g., “non-common lines” must be sufficiently far apart
- Convex relaxation of the log likelihood function using SDP for Unique Games (in progress, with Moses Charikar)
- Translations
- Contrast transfer function of the microscope, different defocus groups
- Colored noise, signal dependent noise
- Beam induced motion
- Heterogeneity

Challenges: What is the resolution?

Put another way, did we get the correct structure?

- No underlying ground truth for comparison, except in simulations (even when a crystal structure is available, it is not necessarily identical to the frozen-hydrated structure)
- Current practice: Fourier Shell Correlation (split data into two halves) (not just a scientific issue – resolution is an NIH criterion for funding)
- Can we estimate bias and variance errors of reconstruction algorithms?
- Analyze refinement procedure (template/reference matching): starting with the ground truth initial model (oracle), or with low-pass filtered ground truth

References

- A. Singer, Y. Shkolnisky, “Three-Dimensional Structure Determination from Common Lines in Cryo-EM by Eigenvectors and Semidefinite Programming”, *SIAM Journal on Imaging Sciences*, **4** (2), pp. 543–572 (2011).
- L. Wang, A. Singer, Z. Wen, “Orientation Determination of Cryo-EM images Using Least Unsquared Deviations”, *arXiv:1211.7045* [cs.LG]
- A. Singer, Z. Zhao, Y. Shkolnisky, R. Hadani, “Viewing Angle Classification of Cryo-Electron Microscopy Images using Eigenvectors”, *SIAM Journal on Imaging Sciences*, **4** (2), pp. 723–759 (2011).
- A. Singer, H.-T. Wu, “Vector diffusion maps and the connection Laplacian”, *Communications on Pure and Applied Mathematics (CPAM)*, **65** (8), pp. 1067–1144 (2012).

Thank You!

Funding:

- NIH/NIGMS R01GM090200
- AFOSR FA9550-12-1-0317
- Sloan Research Foundation
- Simons Foundation LTR DTD 06-05-2012