Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem

Amit Singer

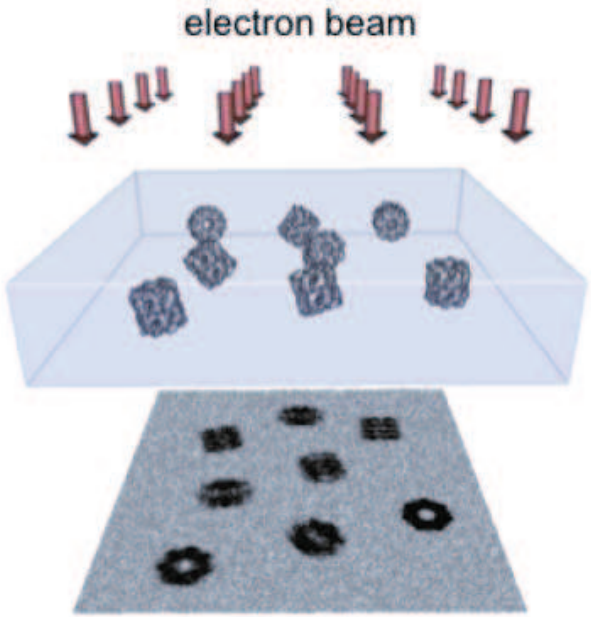
Princeton University, Department of Mathematics and PACM

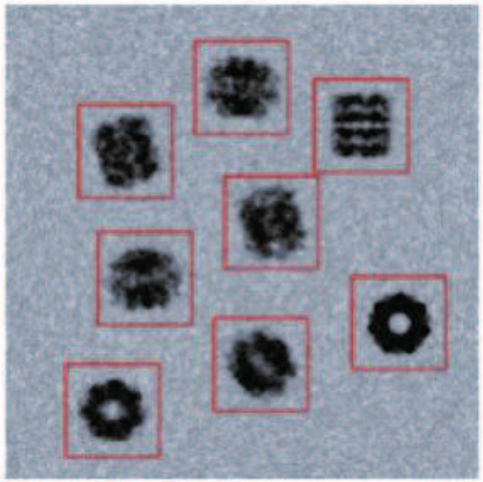
November 11, 2013

Joint work with Gene Katsevich (Princeton) and Alexander Katsevich (UCF)

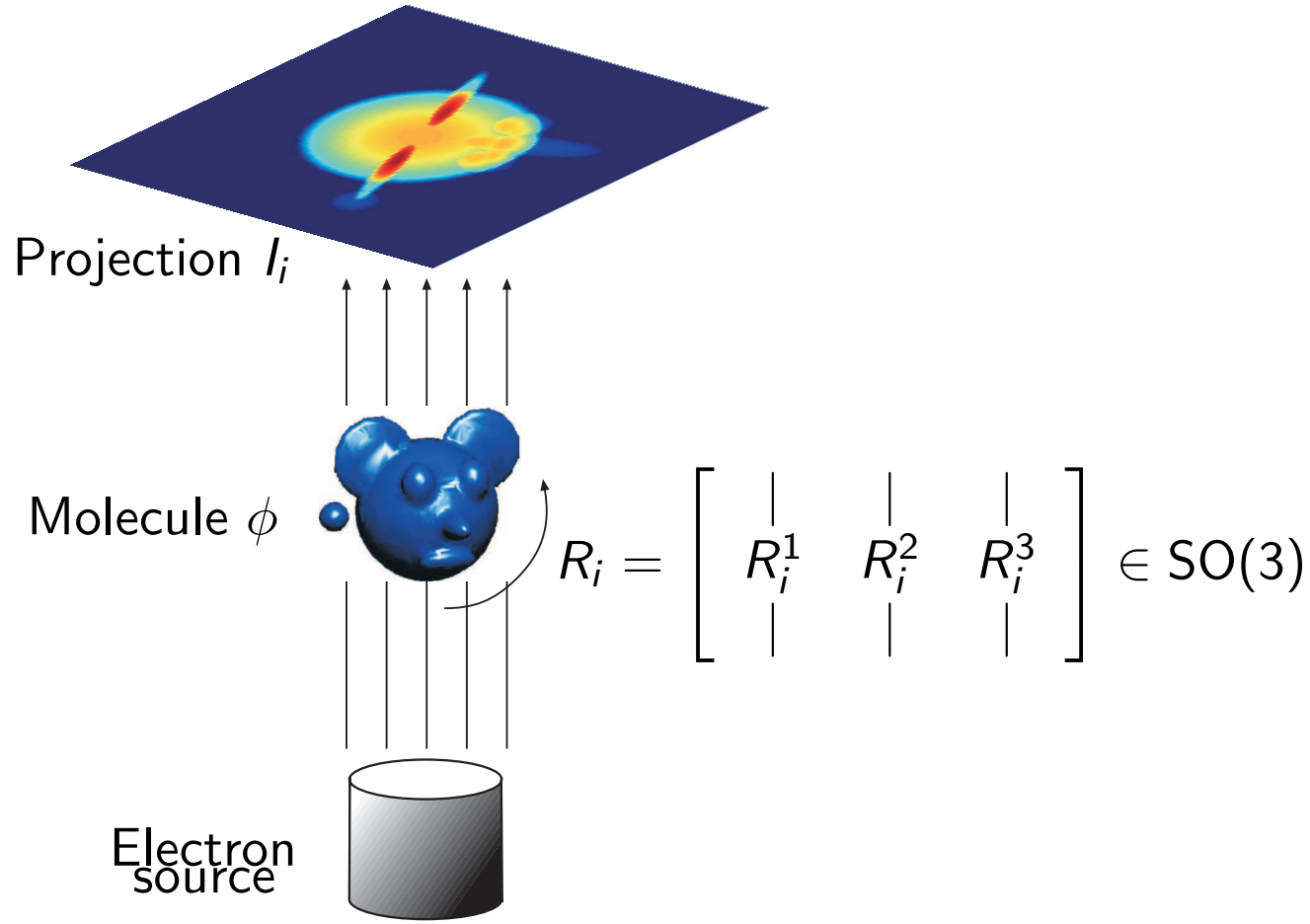
Single Particle Cryo-Electron Microscopy

Drawing of the imaging process:



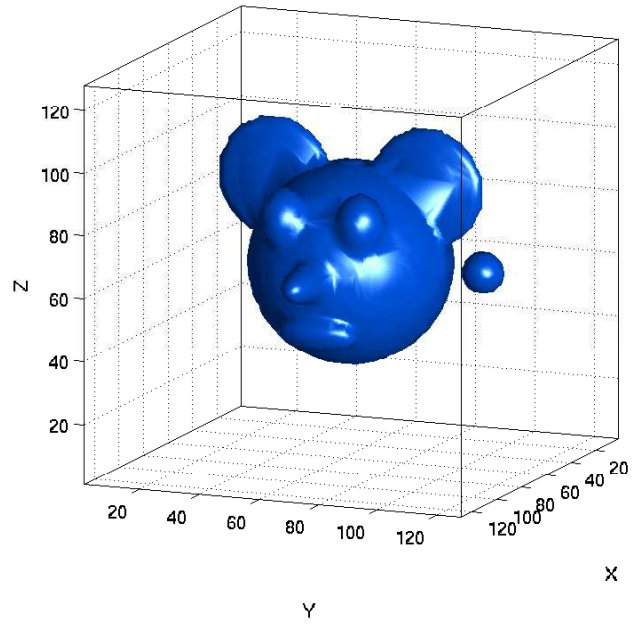


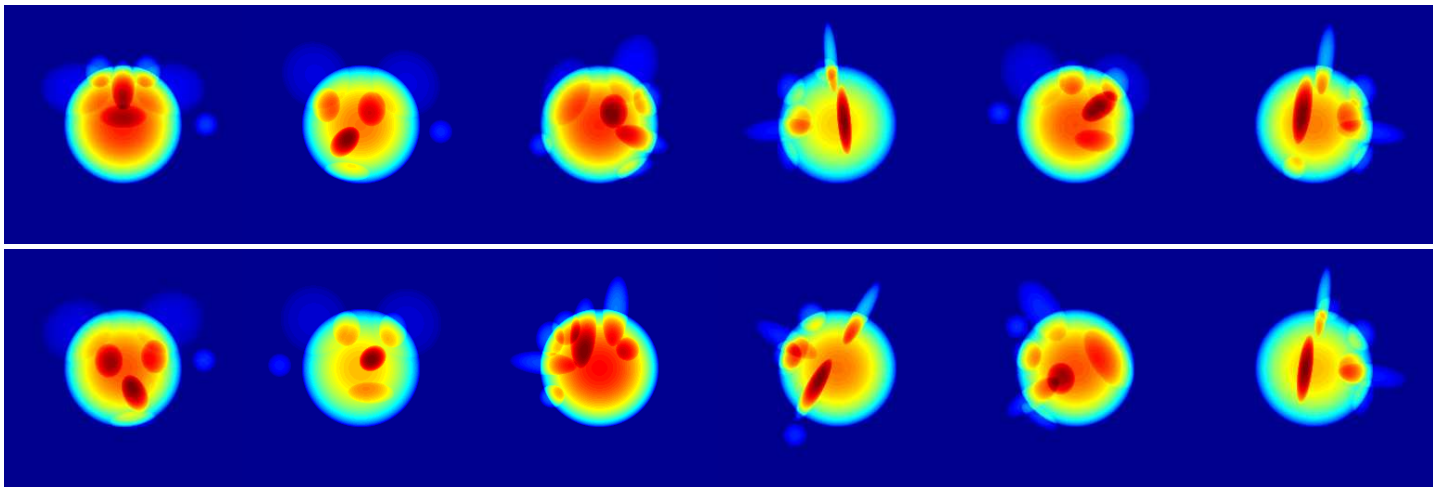
Single Particle Cryo-Electron Microscopy: Model



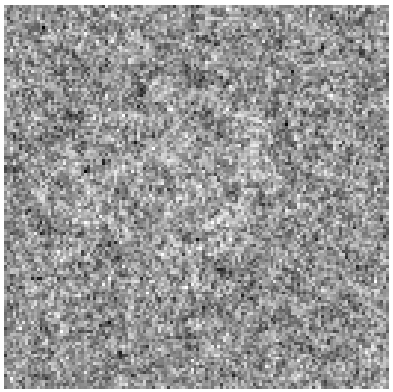
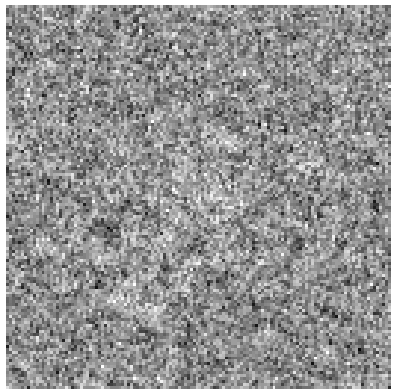
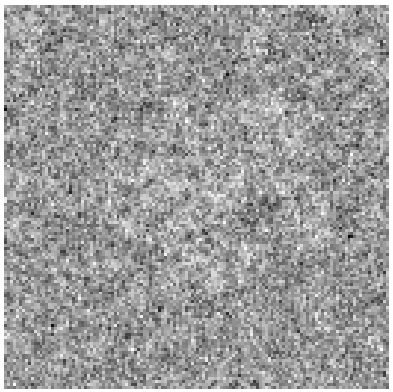
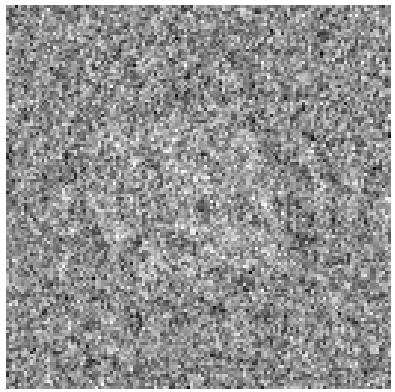
- Projection images $I_i(x, y) = \int_{-\infty}^{\infty} \phi(xR_i^1 + yR_i^2 + zR_i^3)\, dz +$ "noise".
- $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is the electric potential of the molecule.
- Cryo-EM problem: Find ϕ and $R_1, \ldots, R_n$ given $I_1, \ldots, I_n$.

Toy Example





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

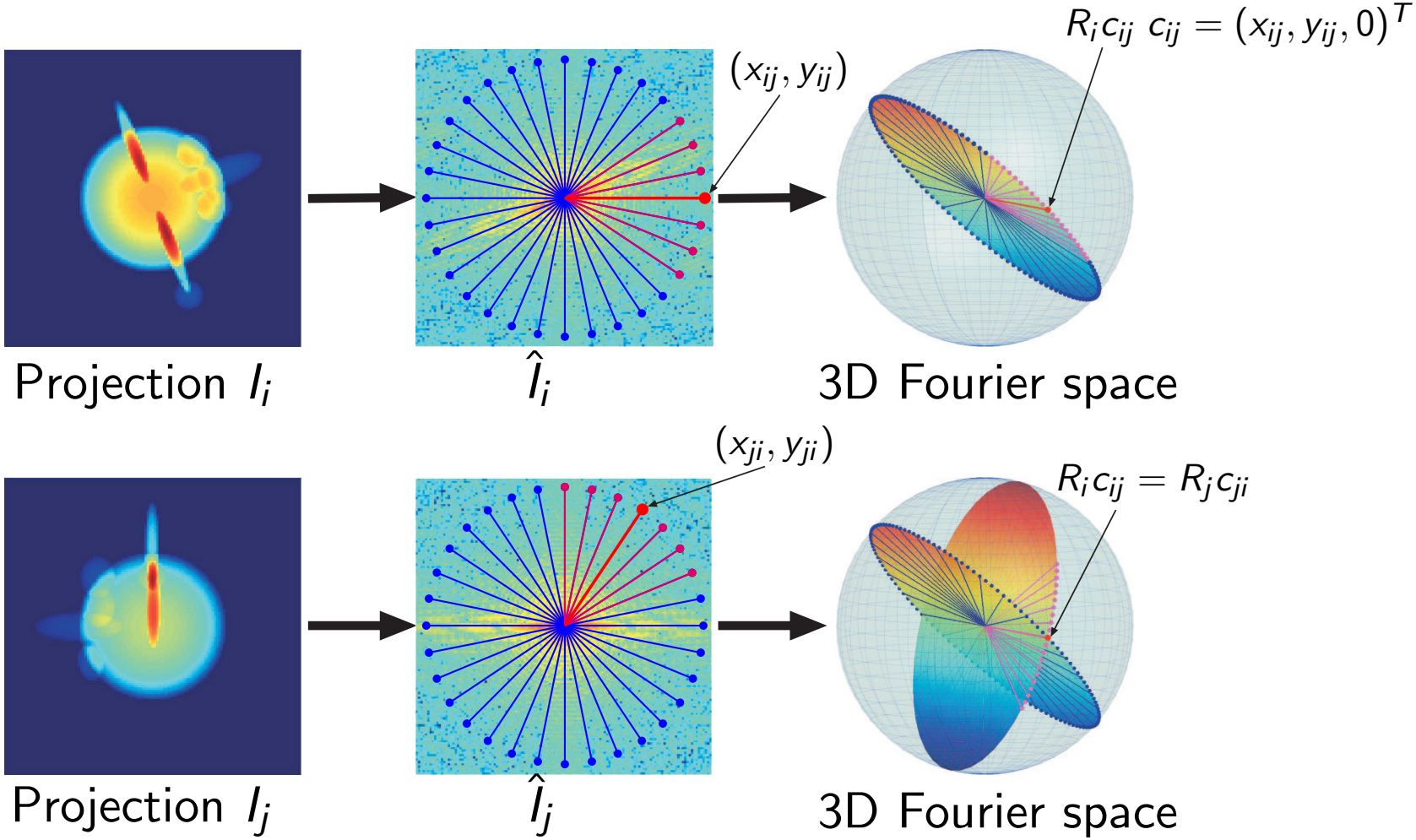





Movie by Lanhui Wang and Zhizhen (Jane) Zhao

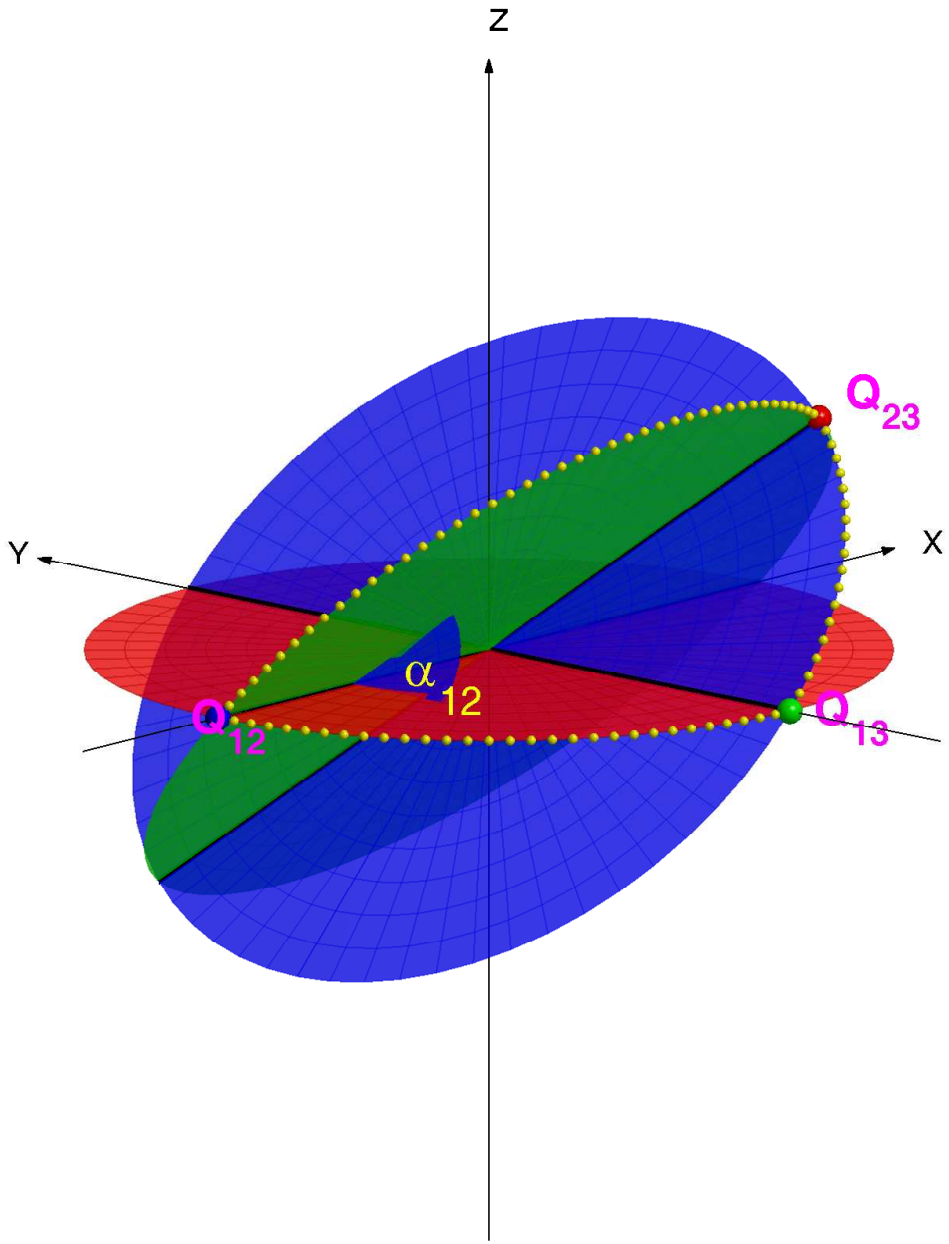
Algorithmic Pipeline

- **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**

Geometry: Fourier projection-slice theorem



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



The Heterogeneity Problem

- A key assumption in classical algorithms for cryo-EM is that the sample consists of (rotated versions of) identical molecules.
- In many datasets this assumption does not hold.
- Some molecules of interest exist in more than one conformational state.
- Examples: A subunit of the molecule might be present or absent, occur in several different arrangements, or be able to move in a continuous fashion from one position to another.
- These structural variations are of great interest to biologists, as they provide insight into the functioning of the molecule.
- Determining the structural variability from a set of cryo-EM images obtained from a mixture of particles of two or more different kinds or different conformations is known as the heterogeneity problem.

The Heterogeneity Problem

- Given 2D projection images of a heterogenous set of 3D volumes, classify the images and reconstruct the 3D volumes.
- One projection image per particle, the projection directions are unknown, and the correspondence between projections and volumes is unknown.
- The underlying distribution of the 3D volumes is unknown: could be a mixture of continuous and discrete, number of classes and/or number of degrees of freedom are also unknown.
- Compared to usual SPR, the effective signal-to-noise ratio (SNR) is even lower, because the signal we seek to reconstruct is the variation of the molecules around their mean, as opposed to the mean volume itself.

Current Approaches

- Penczek et al (JSB 2006): bootstrapping using resampling.
- Scheres et al (Nature Methods 2007): maximum likelihood.
- Shatsky et al (JSB 2010): common lines and spectral clustering.

Do we need more approaches?

While existing methods have their success stories they suffer from certain shortcomings:

- Penczek et al (JSB 2006): bootstrapping using resampling.
  A heuristic sampling method that lacks in theoretical guarantees.
- Scheres et al (Nature Methods 2007): maximum likelihood.
  Requires explicit a-priori distributions, no guarantee for finding global solution, slow (many parameters).
- Shatsky et al (JSB 2010): common lines and spectral clustering.
  Common lines do not exploit all possible information in images.

We would like to have a provable, fast method with low sample complexity that succeeds at low SNR.

Basic Assumption: Small Structural Variability

- We assume that structural variability is small compared to the overall structure.
  For example, variability is confined to a local region.
- Pose parameters of all images are estimated initially as if there is no conformational variability (e.g., using iterative refinement).
- The reconstructed volume is an estimate of the averaged volume (we will address this issue later).
- At this stage, the orientations of all images have been estimated, but classification is still required.
- Our approach would be to perform Principal Component Analysis (PCA) for the 3D volumes given 2D images with known pose parameters.

Principal Component Analysis (PCA)

- PCA is one of the most popular and useful tools in multivariate statistical analysis for dimensionality reduction, compression and de-noising.
- Let $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$ be independent samples of a random vector X with mean and covariance

$$\mathbb{E}[X] = \mu, \quad \mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma$$

- The sample mean and sample covariance matrix are defined as

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_n)(x_i-\mu_n)^T$$

- The principal components are the eigenvectors of Σ_n, ordered by decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$:

$$\Sigma_n v_i = \lambda_i v_i, \quad i = 1, \ldots, p.$$

Classification of 3D Volumes after PCA

- Motivating example: Suppose there are just two dominant conformations, then μ is the average volume and Σ is a rank-1 matrix whose eigenvector is proportional to the difference of the two volumes.
- In general, if there are K classes, then the rank of Σ is at most $K-1$.
- The eigenvectors $v_1, \ldots, v_{K-1}$ are the "eigen-volumes" and enable classification of the projection images.
- If $\phi = \mu + \sum_{k=1}^{K-1} a_k v_k$, then the projection image for rotation R is

$$I_R = P_R \phi + \epsilon = P_R \mu + \sum_{k=1}^{K-1} a_k P_R v_k + \epsilon$$

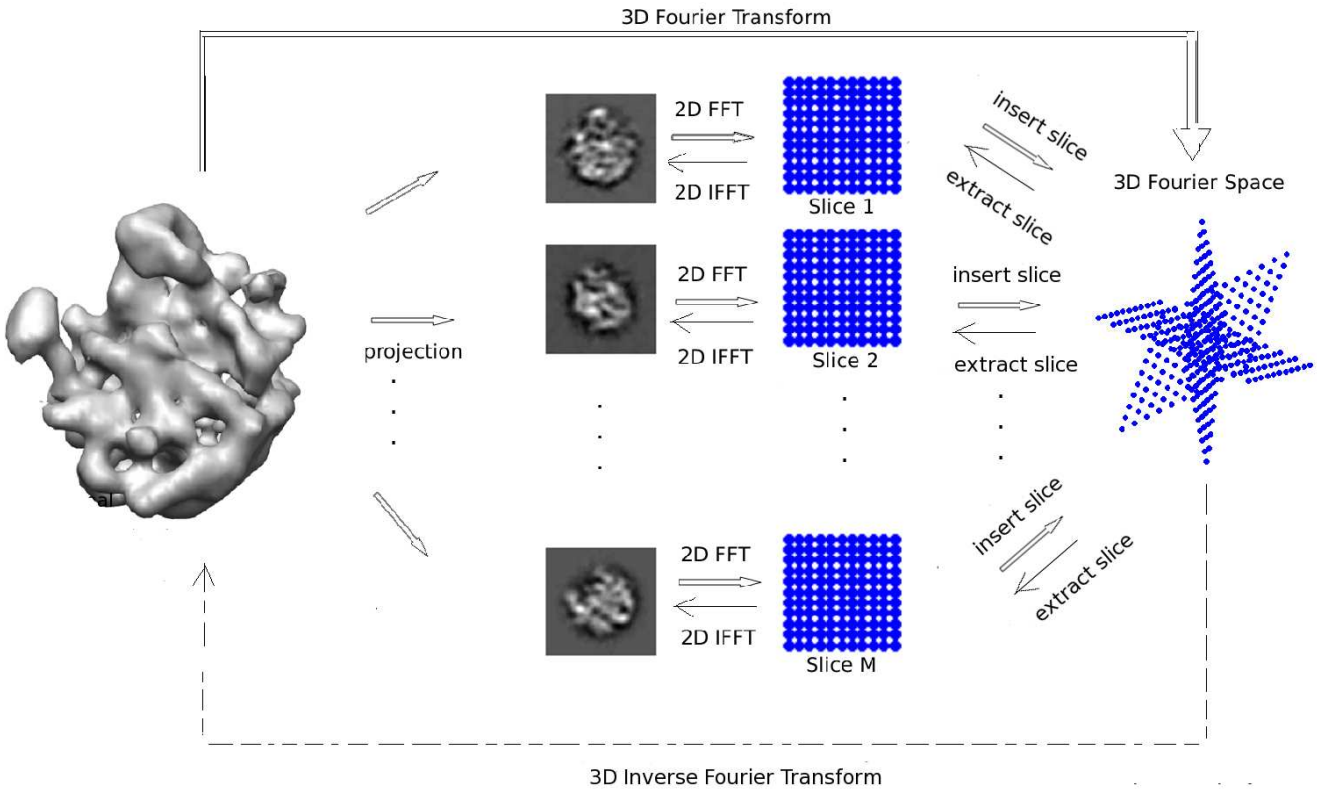
- For each image extract the coefficients $a_1, \ldots, a_{K-1}$ (least squares).
- Use a clustering algorithm (spectral clustering, K-means) to define image classes.

How to estimate the 3D covariance matrix from 2D images?

- In standard PCA, we get samples $x_1, \ldots, x_n$ and we directly construct the sample mean and the sample covariance.
- In the classification problem, the sample mean and sample covariance cannot be computed directly: the covariance matrix of the 3D volumes needs to be estimated from 2D images
- Ad-hoc heuristic solution: Re-sampling — Construct multiple 3D volumes by randomly sampling images and perform PCA for the reconstructed volumes.
- Problems with the resampling approach:
  1. The volumes do not correspond to actual conformations and need not lie on the linear subspace spanned by the conformations
  2. Dependency of volumes due to re-sampling
  3. No theoretical guarantee for accuracy, number of required images, and noise dependency.

Can we estimate the 3D covariance matrix from the 2D images?

- Basic Idea: Fourier projection-slice theorem



Can we estimate the 3D covariance matrix from the 2D images?

- Work in the Fourier domain: It is easier to estimate the covariance matrix of the Fourier transformed volumes
- For any pair of frequencies there is a central slice that contains them. Use all corresponding images to estimate the covariance between those frequencies.
- Repeat to populate the entire covariance matrix.
- If $\hat{\phi} = F\phi$, where F is the 3D DFT matrix, then $\hat{\mu} = F\mu$ and $\hat{\Sigma} = F\Sigma F^*$
- From $\hat{\Sigma}$ we can get Σ. Alternatively, F is a unitary transformation, hence the eigenvectors of Σ and $\hat{\Sigma}$ are related by F.

Limitations of the basic approach - Part I

- Interpolation error: The central slices are sampled on a Cartesian grid that do not coincide with the 3D Cartesian grid. The naïve nearest neighbor interpolation can produce large noticeable errors.

- Statistical error: There are more slices going through some frequencies than others. Examples: low frequency vs. high frequency, frequencies that are on the same central line. Some entries of the covariance matrix are statistically more accurate than others. Classical PCA does not take this into account.

Limitations of the basic approach - Part II

- Sample complexity: How many images are needed as a function of the SNR? How can we tell the "signal" eigenvalues from the "noise" ones? What is the number of groups K? Different from classical PCA since the observations are partial.

- Computational cost: For a volume of size $N \times N \times N = N^3$, the covariance matrix is of size $N^3 \times N^3$. For $N = 100$ this requires 10^{12} storage (4TB), out of RAM. Also computational complexity is nN^4 (where n is the number of images) (typical parameter values $n = 10^5$, $N = 100$ give $nN^4 = 10^{13} = 10$ teraflops).

Research Program

- Interpolation error: We represent the volumes in a spherical-Bessel basis, only using components that satisfy the Nyquist sampling criterion (extension of Klug and Crowther, Nature 1972; Zhao and S, JOSA A 2013).
- Statistical error: We formulate a statistical framework that does not assume an underlying distribution (which is not necessarily Gaussian), and addresses the difficulty imposed by the new basis in which projection operators are no longer "coordinate selection"/restriction.
- Sample complexity: We formulate a certain problem in random matrix theory regarding the limiting spectral density of the eigenvalues of a certain random matrix ensemble.
- Computational cost: We propose different sub-sampling strategies to exploit either the low rank structure of Σ and/or the localized nature of the structural variability.

Mitigating the interpolation error

- Instead of representing a volume using N^3 Cartesian grid voxels, we use a spherical-Bessel expansion:

$$\phi(r,\theta,\varphi) = \sum_{n,l,m} a_{nlm} f_{nlm}(r) Y_l^m(\theta,\varphi)$$

- $f_{nlm}(r) = \frac{J_{l+\frac{1}{2}}(R_{ln} r)}{\sqrt{r}}$ are the spherical-Bessel functions, R_{ln} is the n'th root of $J_{l+\frac{1}{2}}(r) = 0$.
- $Y_l^m(\theta,\varphi)$ are the spherical harmonics.
- $f_{nlm}(r) Y_l^m(\theta,\varphi)$ are the eigenfunctions of the Laplacian in the unit ball with Dirichlet boundary conditions.
- R_{ln}^2 are the eigenvalues of the Laplacian, can be considered as proxy for (squared) frequencies.
- Sampling criterion: keep n, l, m with $l + 2n$ below a certain threshold.
- p is the number of coefficients in the expansion.

Statistical Framework

- We denote the expansion coefficient vectors of the n volumes $\phi_1, \ldots, \phi_n$ by $x_1, \ldots, x_n \in \mathbb{R}^p$.
- These are sampled independently from a distribution over $\mathbb{R}^p$, where $p = O(N^3)$, with

$$\mathbb{E}[X] = \mu, \quad \mathbb{E}[(X - \mu)(X - \mu)^T] = \Sigma.$$

- The distribution is not necessarily Gaussian.
- The linear projection operators $P_1, \ldots, P_n$ from $\mathbb{R}^p$ to $\mathbb{R}^q$ (here $q = N^2$) depend on the rotations and the CTFs.
- Image formation model:

$$I_k = P_k x_k + \epsilon_k, \quad k = 1, \ldots, n,$$

where $\epsilon_k \sim \mathcal{N}(0, \sigma^2 I_{q \times q})$.

- Mean and covariance of an image:

$$\mathbb{E}[I_k] = P_k \mu, \quad \mathbb{E}[(I_k - P_k \mu)(I_k - P_k \mu)^T] = P_k \Sigma P_k^T + \sigma^2 I$$

Statistical Framework

- Define estimators μ_n and Σ_n as minimizers of

$$\mu_n = \underset{\mu}{\operatorname{argmin}} \sum_{k=1}^{n} \| I_k - P_k \mu \|^2$$

$$\Sigma_n = \underset{\Sigma}{\operatorname{argmin}} \sum_{k=1}^{n} \| (I_k - P_k \mu_n)(I_k - P_k \mu_n)^T - (P_k \Sigma P_k^T + \sigma^2 I) \|_F^2$$

- The estimators satisfy

$$\left(\sum_{k=1}^{n} P_k^T P_k \right) \mu_n = \sum_{k=1}^{n} P_k^T I_k$$

$$\sum_{k=1}^{n} P_k^T P_k \Sigma_n P_k^T P_k = \sum_{k=1}^{n} P_k^T [(I_k - P_k \mu_n)(I_k - P_k \mu_n)^T - \sigma^2 I] P_k$$

- μ_n is simply the reconstructed volume in the homogeneous case.
- Notice that Σ_n reduces to the usual sample covariance matrix when no projections are involved.

Statistical Properties of Estimators

The following hold:

- The estimator μ_n is unbiased: $\mathbb{E}[\mu_n] = \mathbb{E}[X]$ (for n large enough so that $\sum_{k=1}^{n} P_k^T P_k$ is full-rank)

- The estimators Σ_n and μ_n are asymptotically consistent:

  $\mu_n \to \mathbb{E}[X]$ and $\Sigma_n \to \mathrm{Cov}(X)$ almost surely as $n \to \infty$.

Sample complexity requires further investigation, since in practice we are outside the realm of classical statistics ($n \gg p$), but rather in our case $n \approx p$.

"High-dimensional" statistics.

The operator L

$$\sum_{k=1}^{n} P_k^T P_k \Sigma_n P_k^T P_k = \sum_{k=1}^{n} P_k^T [(I_k - P_k \mu_n)(I_k - P_k \mu_n)^T - \sigma^2 I] P_k$$

- Σ_n requires the inversion of the linear operator L (a matrix of size $p^2 \times p^2$):

$$L(\Sigma) = \sum_{k=1}^{n} P_k^T P_k \Sigma P_k^T P_k$$

- We devised a fast algorithm to invert L (notice that the inversion is fast when P_k are "coordinate selection" operators but not necessarily so in general).
- The fast algorithm is based on the continuum limit of L

$$\mathcal{L}(\Sigma) = \int_{SO(3)} P(R)^T P(R) \Sigma P(R)^T P(R) \, d\mu(R)$$

where integration is with respect to the Haar measure over $SO(3)$.

Projection covariance transform

$$\mathcal{L}(\Sigma) = \int_{SO(3)} P(R)^T P(R) \Sigma P(R)^T P(R) \, d\mu(R)$$

- $\mathcal{L}$ is positive semidefinite and commutes with rotations, and has a block-diagonal sparse structure.
- We call $\mathcal{L}$ the **projection covariance transform**.
- $\mathcal{L}$ is important for covariance estimation for inverse problems involving structural variation just as projection and back-projection operators are important for classical inversion problems.

Sample complexity – a problem in random matrix theory

- Recall Σ_n is defined through

$$L(\Sigma_n) = \sum_{k=1}^{n} P_k^T [(I_k - P_k \mu_n)(I_k - P_k \mu_n)^T - \sigma^2 I] P_k$$

- First need to understand the spectrum (eigenvalues) due to pure noise (i.e., $\mu = \mu_n = 0$, $\Sigma = 0$) for the random matrix in the rhs:

$$S_n = \frac{1}{n} \sum_{k=1}^{n} P_k^T \epsilon_k \epsilon_k^T P_k$$

- S_n is the sample covariance of $y_k = P_k^T \epsilon_k$.
- $\mathbb{E}[y_k] = 0$ and $\mathbb{E}[y_k y_k^T] = \sigma^2 \mathbb{E}[P^T P]$.
- $\mathbb{E}[P^T P]$ is a classical operator in tomography, eigenvalues can be computed explicitly.
- Marcenko and Pastur (1967) derived the spectral density (via the Stieltjes transform) for non-isotropic distributions in the limit $p/n \to \gamma$.

Sample complexity – a problem in random matrix theory

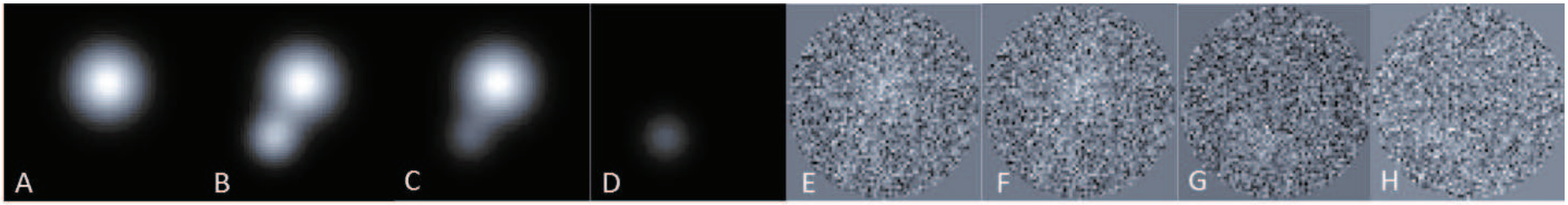
- Recall Σ_n is defined through

$$L(\Sigma_n) = \sum_{k=1}^{n} P_k^T [(I_k - P_k \mu_n)(I_k - P_k \mu_n)^T - \sigma^2 I] P_k$$

$$S_n = \frac{1}{n} \sum_{k=1}^{n} P_k^T \epsilon_k \epsilon_k^T P_k$$

- The effect of inverting L is however less understood.
- What is the limiting spectral density of $L^{-1}(S_n)$?
- What is the distribution of the largest eigenvalue of $L^{-1}(S_n)$? Important for determining the number of heterogeneous groups that can be inferred reliably from the data.

Numerical results



A: projection of the first phantom $e^{-\frac{1}{100}\|r-c_1\|^2}$

B: projection of the second phantom $e^{-\frac{1}{100}\|r-c_1\|^2} + e^{-\frac{1}{50}\|r-c_2\|^2}$

C: mean projection

D: deviation from mean

E: noisy projection SNR=0.05

F: noisy projection SNR=0.01

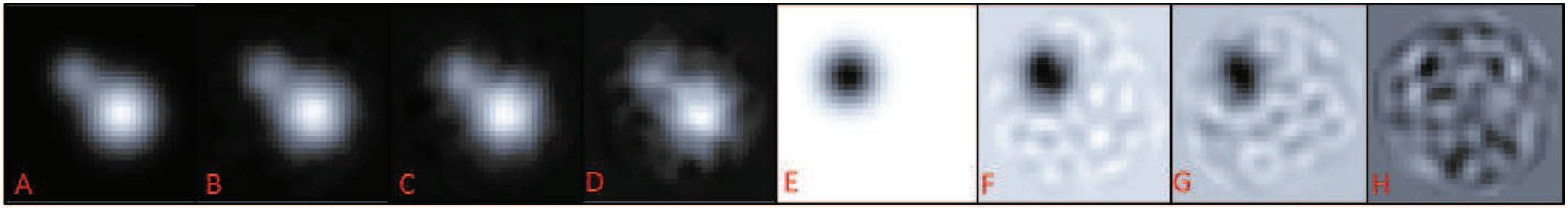
G: after subtraction of mean SNR=0.05

H: after subtraction of mean SNR=0.01.

Observe that the signal is much weaker in G,H compared to E,F.

Numerical results

$n = 10000$ projection images



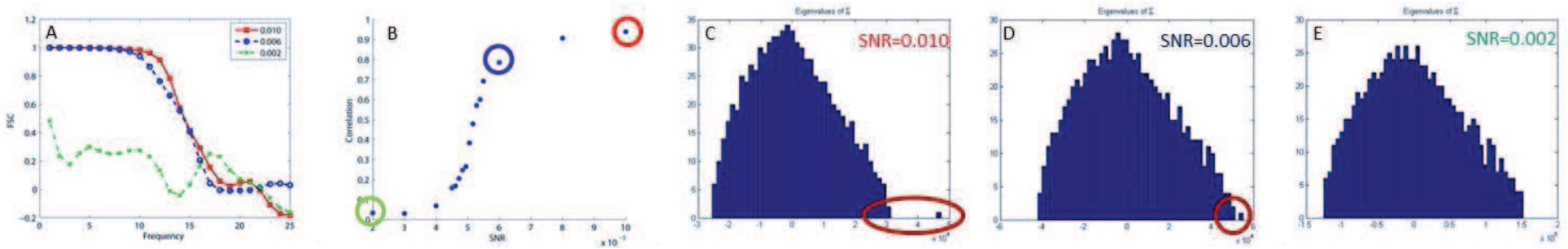
A: slice of the true mean

B, C, D: slices of reconstructed mean SNR=0.01, 0.006, 0.002 (resp.)

E: slice of true volume difference

F, G, H: slices of reconstructed leading eigenvector SNR=0.01, 0.006, 0.002 (resp.)

Numerical results



A: Fourier Shell Correlation (FSC) curves for the top eigenvector at the same three SNRs as before

B: normalized cross-correlation of the computed top eigenvector with its true value for different SNRs

C, D, E: eigenvalue histograms of reconstructed covariance matrix for three SNR values. Note that the noise distribution comes increasingly closer to the top eigenvalue, and eventually the latter is no longer distinguishable.

The correlation values in A and B depend on the size of the spectral gap in C, D, E.

Summary

- PCA is a viable method for tackling the heterogeneity problem.
- It is possible to estimate the 3D covariance matrix directly from 2D projection images accurately and efficiently.
- No need to resort to heuristic approaches such as bootstrapping using resampling.
- No need to impose a prior on the distribution of conformations.
- Random matrices play an important role in solving the heterogeneity problem.

Thank You!

Funding:

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

G. Katsevich, A. Katsevich, A. Singer (submitted). Covariance Matrix Estimation for the Cryo-EM Heterogeneity Problem. http://arxiv.org/abs/1309.1737