3D Classification

Marin van Heel
Leiden University/Imperial College London/LNNano Campinas

Strasbourg 7 October 2016
Fundamentals operations in reconstructions and projections

Real Space

3D-structure ("map")

project map

back project image

2D projection image

project to line

back project line

1D (line) projection (sinogram line)

Fourier Space

3D-FT

extract centr. sect.

insert centr. sect.

2D-FT (central section)

extract centr. line

insert centr. line

1D-FT (central line)

(van Heel 1987)
Fundamentals operations with heterogeneous data

Real Space

- 3D-structure ("map")
  - 3D-FT
  - 2D-FFT
  - 1D-FFT
  - Project map
  - Project image
  - Project line (sinogram line)

Fourier Space

- 3D-structure ("map")
  - 3D-FT
  - 2D-FT (central section)
  - 1D-FT (central line)
  - Extract central section
  - Insert central section
  - Insert central line
  - Extract central line

Van Heel 1987
3D Classification

Marin van Heel
Leiden University/Imperial College London/LNNano Campinas

Strasbourg 7 October 2016
Classification ...?

Supervised Classification

(looking for a specific thing, matched filtering, find my reference, reference bias)

Unsupervised Classification

(comparing everything to everything and let the data speak for itself, avoid bias)
Classification ...?

Nomenclature has become a mess...

2D Classification?
3D Classification?
A priori frustration:

(you decide what makes sense, not the classification program)
Classification is a matter of perspective...

- Small fish: There is no justice in the world.
- Average fish: There is some justice in the world.
- Large fish: The world is just.

(MANKOFF)
Biology is a mess...

(you need powerful multivariate statistical tools to make sense of it all)

Multivariate statistics:

is all about distances, correlations...

Correlation / Inner Product

\[ = \sum_a F(a) \cdot G(a) \]

(Euclidian Distance)^2 = \[ = \sum_a (F(a) - G(a))^2 \]

Dist. \[^2 = \sum_a \left( F^2(a) - 2(F(a) \cdot G(a)) + G^2(a) \right) \]

Dist. \[^2 = \sum_a F^2(a) - 2 \sum_a (F(a) \cdot G(a)) + \sum_a G^2(a) \]

Normalised Signal = F(a) / \[ \sqrt{\sum_a F^2(a)} \] = F(a) / SD_F
Hyperspace Data Representation of single particle images

Data Compression!
Hyperspace representation of single-particle data

Data Compression!

J-P Bretaudière
Hyperspace representation and Classification
Data Matrix "X"

\[ X = \begin{pmatrix}
  x_{1,1} & x_{1,2} & x_{1,3} & \cdots & \cdots & \cdots & x_{1,p} \\
  x_{2,1} & x_{2,2} & x_{2,3} & \cdots & \cdots & \cdots & x_{2,p} \\
  \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
  x_{n,1} & x_{n,2} & x_{n,3} & \cdots & \cdots & \cdots & x_{n,p}
\end{pmatrix} \]
Eigenimages-i"}

\[ E_1' \]
\[ E_2' \]

* 

\[ \text{Aligned Images} \]

\[ I_1 \]
\[ I_2 \]
\[ I_3 \]

Ortho-norm. + Over-relax.

\[ \text{Inner Products} \]

\[
\begin{array}{c|c}
C_{11} & C_{12} \\
C_{21} & C_{22} \\
\vdots & \vdots
\end{array}
\]

\[ C_{ji} = I_j \cdot E_i' \]

\[ \text{Eigenimages-i} \]

\[ E_1 \]
\[ E_2 \]

\[ \sum C_{ji} \cdot I_j \]

(Ref.: Review 2000)
From algorithm to mathematics:

$$X' \cdot X \cdot U = U \cdot \Lambda$$

In detail eigenvector equation:

$$X' \cdot N \cdot X \cdot M \cdot U = U \cdot \Lambda$$

With orthonormalisation constraint:

$$U' \cdot M \cdot U = I_p$$

In conjugate space:

$$X \cdot M \cdot X' \cdot N \cdot V = V \cdot \Lambda$$

With orthonormalisation constraint:

$$U' \cdot M \cdot U = I_p$$
Parallel MSA and its scaling with the number of available CPUs (tested on 32 nodes with each 4 CPUs). The calculations necessary for standard MSA algorithm (top left) are distributed over the available CPUs (lower left). I/O is parallelized by copying the relevant part of the huge input data file to the local scratch file available on each node. Overall speed increase with the current version of the MSA (see text) and images of size 256x256 is around 27 times (Full Program).
MSA analysis of *Lumbricus Terrestris* hemoglobin
Automatic Classification

Average images in each class
Fig. 1. An example of a hierarchical classification tree. In an hierarchical ascendant classification the procedure starts at the bottom of the figure with as many classes as there are images (10 in this example). The two classes that are closest together in terms of a classification criterion are merged into a larger class. The straight cut through the classification tree leads to a variance-oriented partitioning. A useful alternative is to follow the tree up and down, to obtain a class-size-oriented partition. For details see text.
Ward Criterion

\[ \text{Add. Var.}_{i,i'} = \frac{w_i w_i'}{w_i + w_i'} d_{i,i'}^2 \]
ABC-4D (Alignment by classification)

1000 class averages (144,000 particles)
From 2D classes to 3D Structure(s)
Three-Dimensional reconstruction from projections

We first need to find the relative angles to do a 3D reconstruction…
Intersecting Central Sections
(DeRosier Klug 1968; Crowther 1971)
Fundamentals operations in 3D reconstructions and projections

Real Space

- 3D-structure ("map")
  - project map
  - back project image

- 2D projection image
  - project to line
  - back project line

Fourier Space

- 3D-FT
  - 3D-FFT

- 2D-FT (central section)
  - extract centr. sect.
  - insert centr. sect.

- 1D-FT (central line)
  - extract centr. line
  - insert centr. line

1D (line) projection (sinogram line)

(van Heel 1987)
Fundamentals operations in reconstructions and projections

Projection Matching →

Real Space
- 3D-structure ("map")
  - project map
  - back project image
- 2D projection image
  - project to line
  - back project line
- 1D (line) projection (sinogram line)

Fourier Space
- 3D-FT
  - 3D-FFT
- 2D-FT (central section)
  - extract centr. sect.
  - insert centr. sect.
- 1D-FT (central line)
  - extract centr. line
  - insert centr. line

Angular → Reconstitution

Central Section ← Matching

← Fourier space

Common Lines
Width of Central Section: Central Section Slab

**Reciprocity Real and Fourier Space**

**Real Space**

**Projection Direction**

- $D$
- $L$

**Fourier Space**

- $1/L$
- $1/D$
- $1/D$
- $1/D$
- $1/D$
- $1/D$
- $1/D$
- $1/D$
Fig. 14. Overlapping central sections. Fourier-space central sections, associated with 2D projections a 3D object of linear size ‘$D$’, have a width ‘$1/D$’. Central sections in Fourier space always overlap at very low frequencies, that is, close to the origin. Neighbouring central section, separated by an angle $\phi$, cover largely the same information up to spatial frequency ‘$f_c$’. The overlap of central sections is fundamental in both 3D reconstruction algorithms and in determining the highest isotropic resolution achievable for a given 3D reconstruction geometry. The areas of the 3D Fourier space volume not covered by the central section ‘slabs’ are not measured and are referred to as ‘missing cone’ or ‘missing wedge’ depending on the 3D reconstruction geometry.
Early projection matching experiments: Budapest 1984 Abstract

Figure 1. Continuous stereographic representation of "phantom" used to investigate properties of self-optimizing 3D reconstruction.

Figure 2. Stereo representation of "phantom" reconstructed using random starting angles; viewed from same directions as above.
Angular Reconstitution

(MvH 1987)
70S *E. coli* projection plus sinograms
Intersecting Central Sections in 3D Fourier space

We have only measured the information inside the central section slabs, not the information in the missing wedges.
4D Data Processing
RF3 Complex
Type-1
vs.
Type-2


4D cryo-EM!
IMAGIC 4D:

A priori 3D assignment

3D-1

3D-2

3D-3

3D rec.

Competitive 3D assignment

3D-1'

3D-2'

3D-3'

3D rec.

Update 3Ds
A typical 4D refinement round

<table>
<thead>
<tr>
<th>Histogram of OLD 3D references</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
</tr>
<tr>
<td>312 1</td>
</tr>
<tr>
<td>293 2</td>
</tr>
<tr>
<td>188 3</td>
</tr>
<tr>
<td>587 4</td>
</tr>
<tr>
<td>0 5</td>
</tr>
</tbody>
</table>

Meaning of columns: # in this bin, bin value

<table>
<thead>
<tr>
<th>Histogram of NEW 3D references</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
</tr>
<tr>
<td>278 1</td>
</tr>
<tr>
<td>304 2</td>
</tr>
<tr>
<td>175 3</td>
</tr>
<tr>
<td>623 4</td>
</tr>
<tr>
<td>0 5</td>
</tr>
</tbody>
</table>

Meaning of columns: # in this bin, bin value
A typical 4D refinement round

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1380</td>
<td>0</td>
<td>0</td>
<td>4.78E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1345</td>
<td>35</td>
<td>35</td>
<td>5.81E+01</td>
<td>****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td>360</td>
<td>325</td>
<td>6.33E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>641</td>
<td>739</td>
<td>379</td>
<td>6.85E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>436</td>
<td>944</td>
<td>205</td>
<td>7.37E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>341</td>
<td>1039</td>
<td>95</td>
<td>7.89E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>1124</td>
<td>85</td>
<td>8.41E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td>1174</td>
<td>50</td>
<td>8.93E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>1217</td>
<td>43</td>
<td>9.45E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>1261</td>
<td>44</td>
<td>9.96E+01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>1291</td>
<td>30</td>
<td>1.05E+02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>1310</td>
<td>19</td>
<td>1.10E+02</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>1321</td>
<td>11</td>
<td>1.15E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>1338</td>
<td>17</td>
<td>1.20E+02</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1348</td>
<td>10</td>
<td>1.26E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1355</td>
<td>7</td>
<td>1.31E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1361</td>
<td>6</td>
<td>1.36E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1363</td>
<td>2</td>
<td>1.41E+02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1368</td>
<td>5</td>
<td>1.46E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1380</td>
<td>12</td>
<td>1.57E+02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Low values binned at lower edge: 0
High values binned at higher edge: 12

Meaning of columns: # remaining, # accumulated, # in this bin, bin value
Stark et al. Nature 2010
RF3 Complex Type-1 vs. Type-2 (Klaholz et al., Nature, 2004)

4D cryo-EM!

Focussed classification in 2D
Focussed classification in 3D/4D: process each Globin fold independently
Pavel Afanasyev (Leiden/Maastricht)
Charlotte Linnemayr-Seer (Zurich)
Bart Alewijnse (NeCEN)
Michael Schatz (ImSc/Berlin)
Ralf Schmidt (ImSc/Berlin)
Sacha de Carlo (FEI/NeCEN)
Rishi Matadeen (NeCEN)
Rodrigo Portugal (LNNano)

and many others…