Spectral clustering for microarray data

Alvin Liang

Grant Wang

Cameron Wheeler

12 May 2005

Outline

- Introduction to spectral algorithm
- Application to Golub et al. leukemia data set
- Application to Devauchelle et al. arthritis data set

Introduction to the spectral algorithm

- Two steps:
  - Create hierarchical tree in “top-down” manner based on eigenvalues, eigenvectors
  - Merges leaves in “bottom-up” fashion to create clusters based on objective function
    - Informally, objective function maximizes:

      (“similarity” inside clusters) - (“similarity” outside clusters)
    - You choose what “similarity” means
- Algorithm has been applied in other contexts (web search)

Golub et al. data

- 38 “training” patients, 34 “test” patients
- Cluster into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML)
- They consider “class discovery”: What if you didn’t know about the ALL/AML distinction? How would you find it?
- Their answer: clustering

Their results

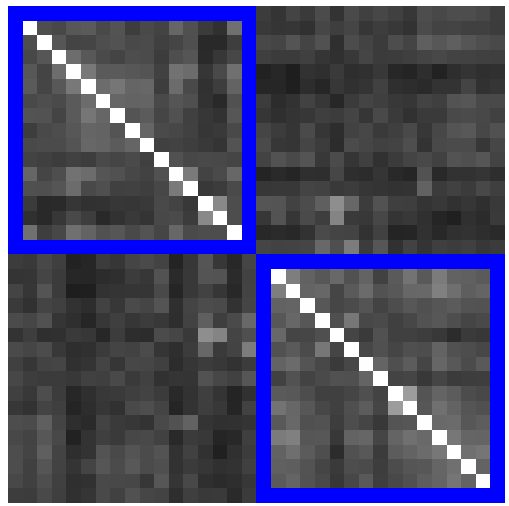
- Only 4 errors out of 34 using self-organizing maps, in two clusters (they used GENECLUSTER)
- How does spectral clustering do?
- How do we present the data to the spectral clustering algorithm?

Normalization

- Spectral clustering wants similarity between patients (between 0,1)
- How do we do it?
  - Normalize data so mean = 0, variance =1
    - (same as Golub et al)
  - Put all positive genes in one coordinate
  - Put all negative genes in another
- Now dot product between two patients is similarity

Our results

- Creates one cluster with (14 ALL, 3 AML)
- Creates another cluster with (8 ALL, 9 AML)
- Why doesn't it work?
- Let's look at three different clusterings:
  - "Correct" clustering (all ALL in one, all AML in other
  - "Our clustering"
  - Random clustering



What is a good clustering?

- If a clustering of patients is good, then there should be genes that are expressed high in that cluster, and low in the other.
  - (I.e. genes that “differentiate” the two clusters)
- We find:
  - “Correct clustering”: yes, there are such genes
  - Our clustering: yes, there are such genes
  - Random clustering: no, no such genes

Results

- True clustering:

ALL	AML	diff	
0.00	0.92	-0.92	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
0.05	0.92	-0.87	APLP2 Amyloid beta (A4) precursor-like protein 2
0.00	0.85	-0.85	CD33 CD33 antigen (differentiation antigen)

- Our clustering:

C1	C2	diff	
0.11	1.00	-0.88	GB DEF = Secreted epithelial tumor mucin antigen
0.17	1.00	-0.82	KIAA0265 gene, partial cds
0.11	0.94	-0.82	Hyaluronoglucosaminidase 1 (HYAL1) mRNA

- Random clustering:

C1	C2	diff	
0.2353	0.8235	-0.5882	n/a
0.1765	0.7059	-0.5294	n/a
0.1765	0.7059	-0.5294	n/a

OA vs RA

Four images removed for copyright reasons.

Seek to Classify OA and RA

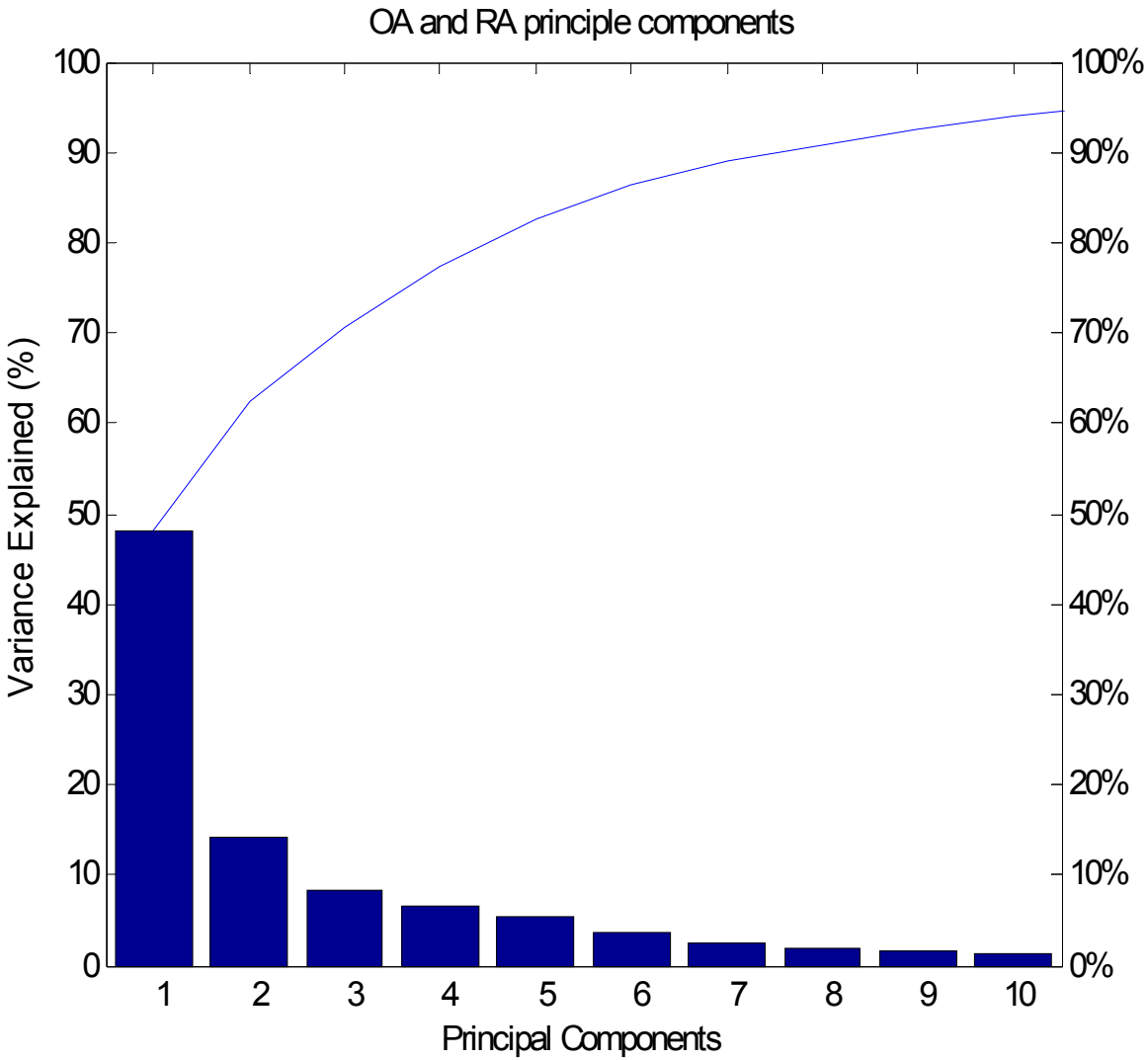
- Use Traditional Clustering Methods
  - Classification
  - Gene Clustering
- Spectral Clustering
  - Classification
  - Gene Clustering

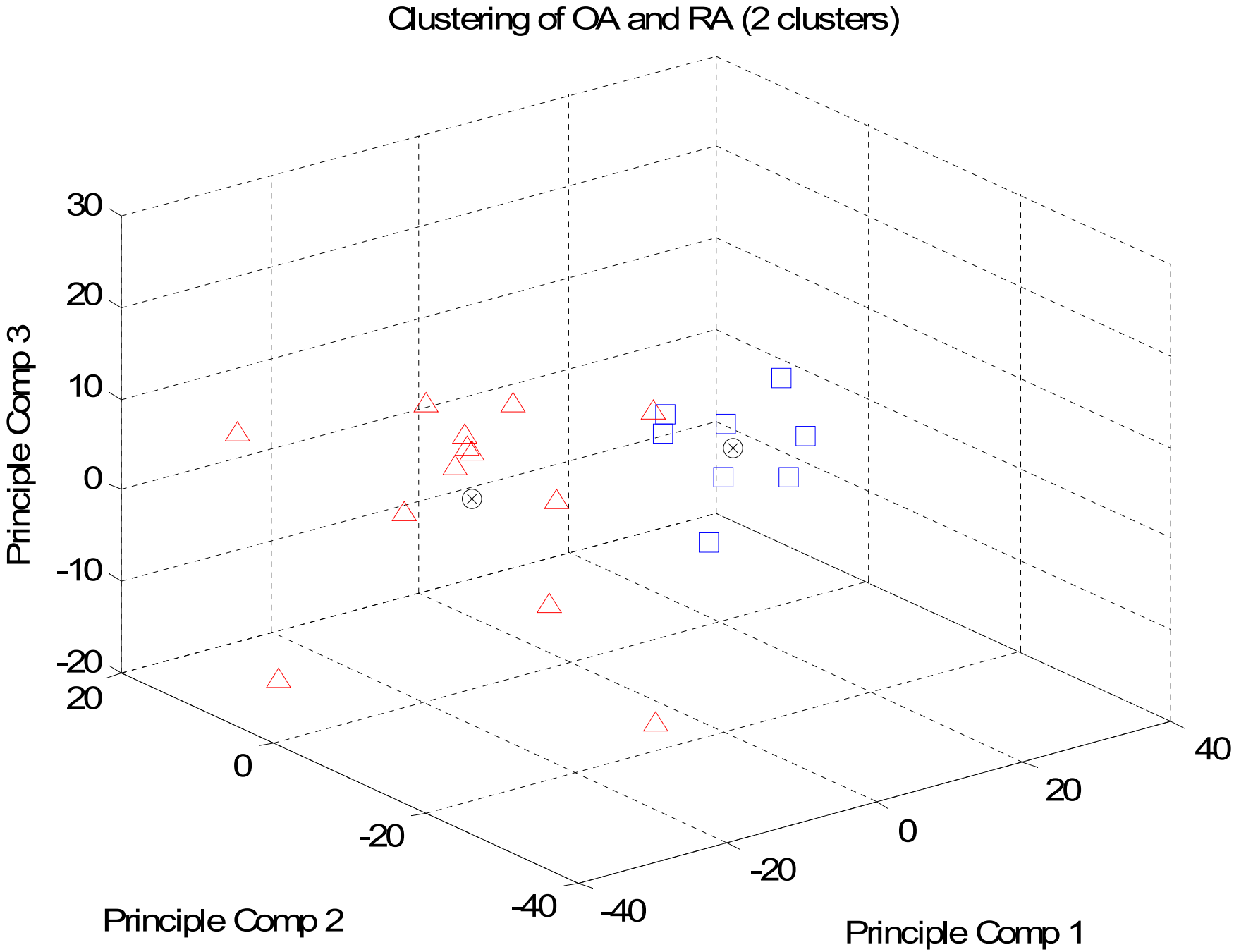
Data from Devauchelle *et al*.

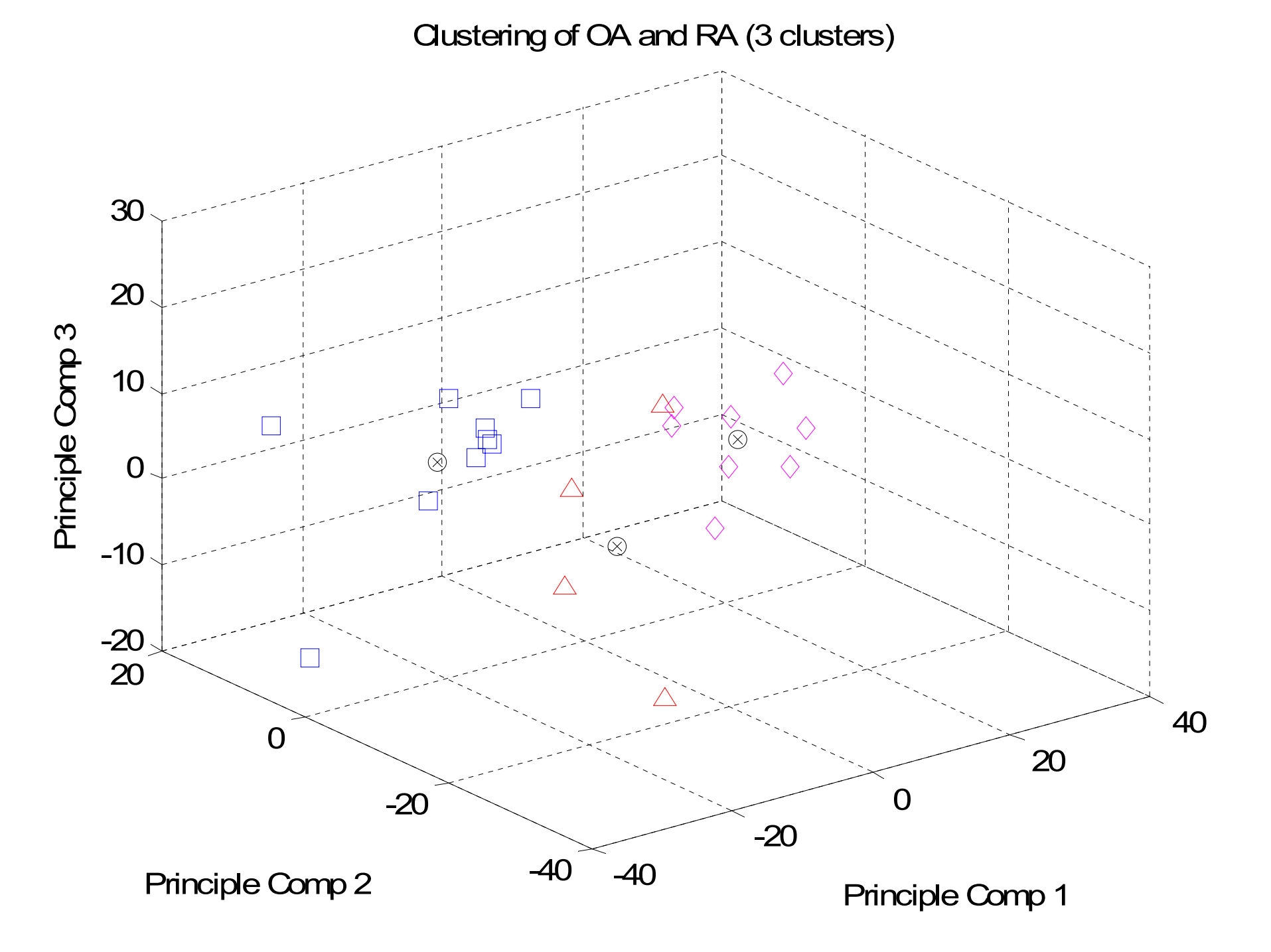
- Synovial tissue
  - 13 OA patients (age 72 ± 9.3)
  - 8 RA patients (age 55± 9.2)

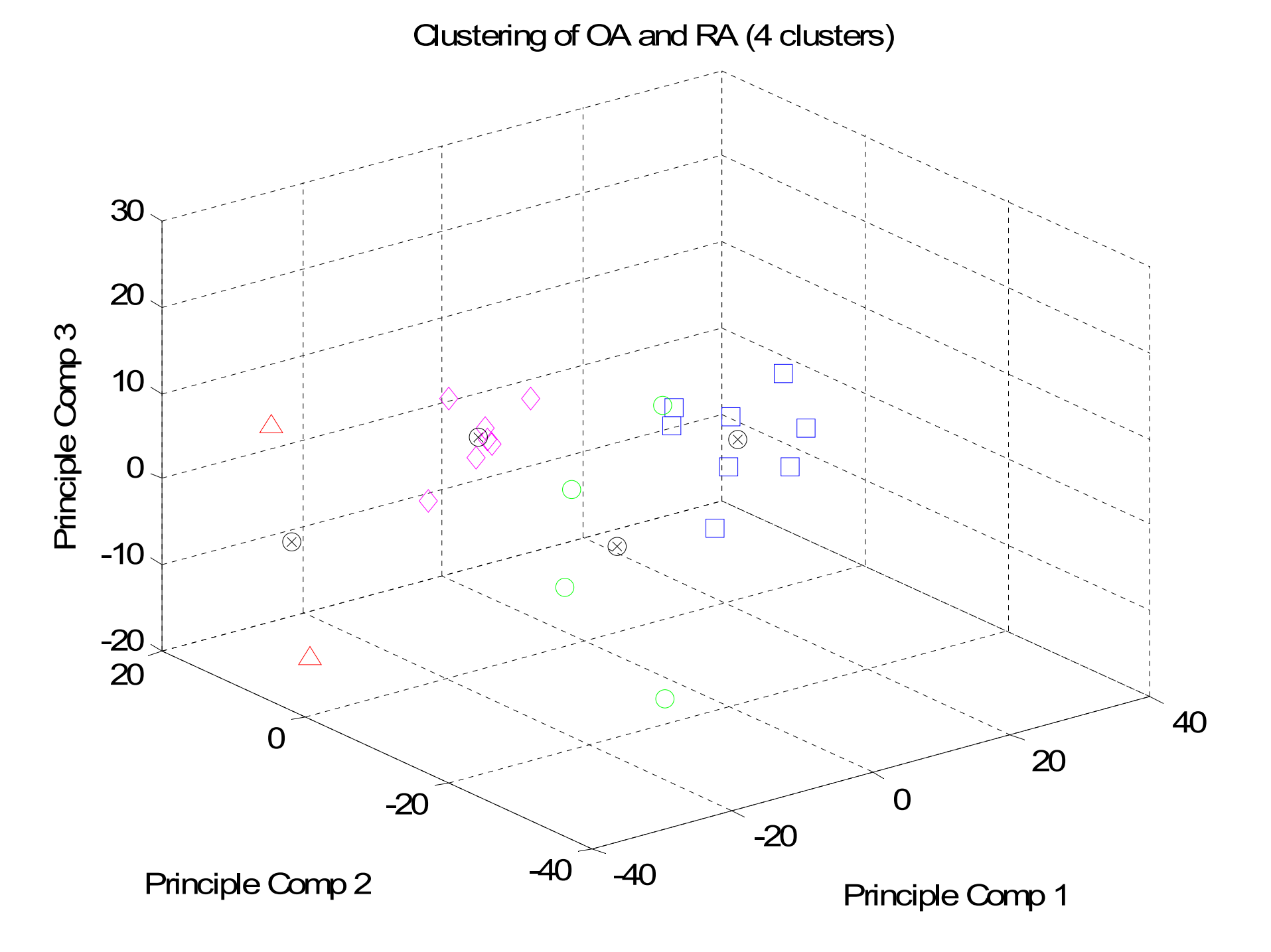
- 4652 Genes probed
  - 63 genes selected
  - Normalized to mRNA levels

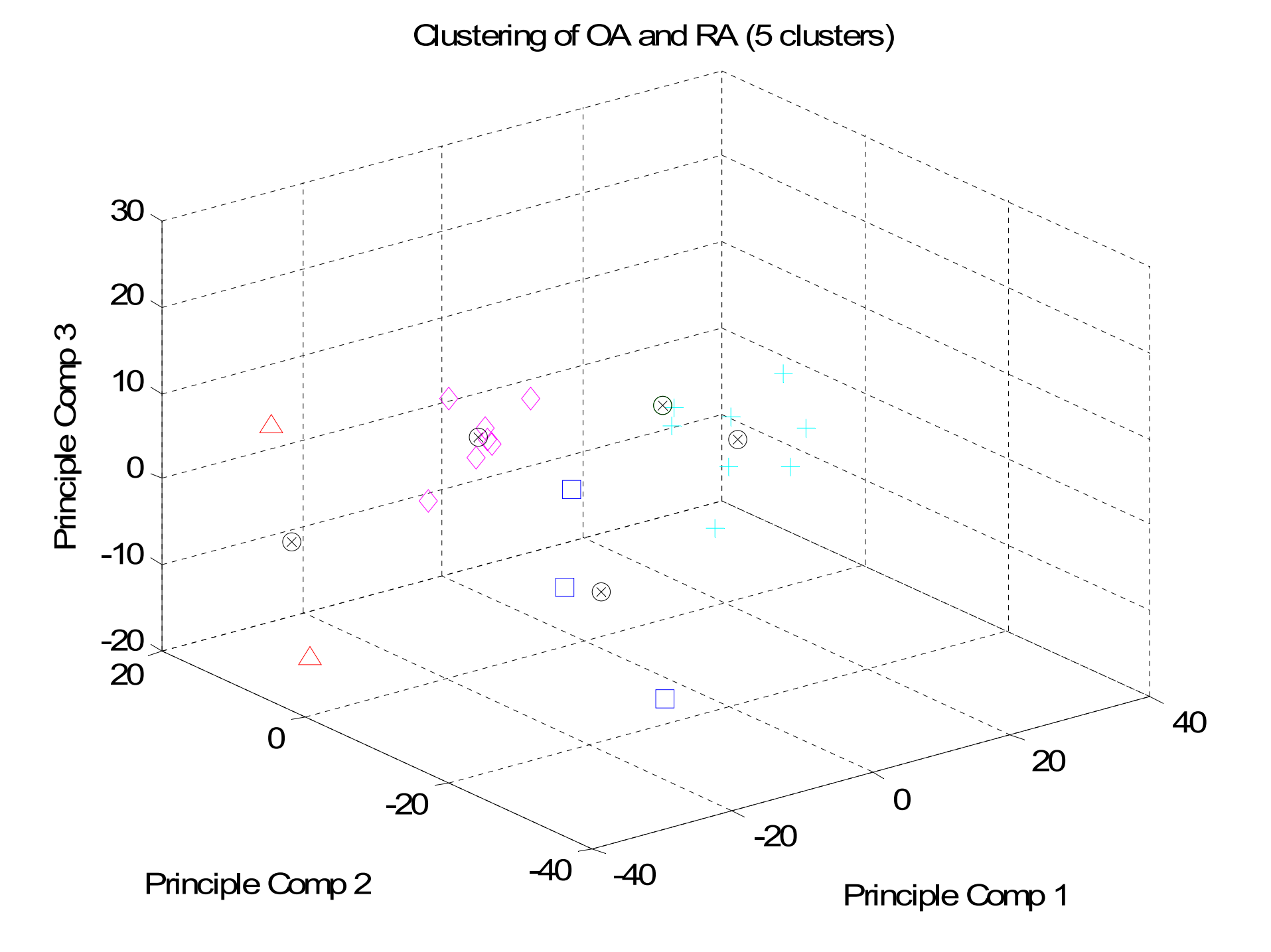
PCA and K-means



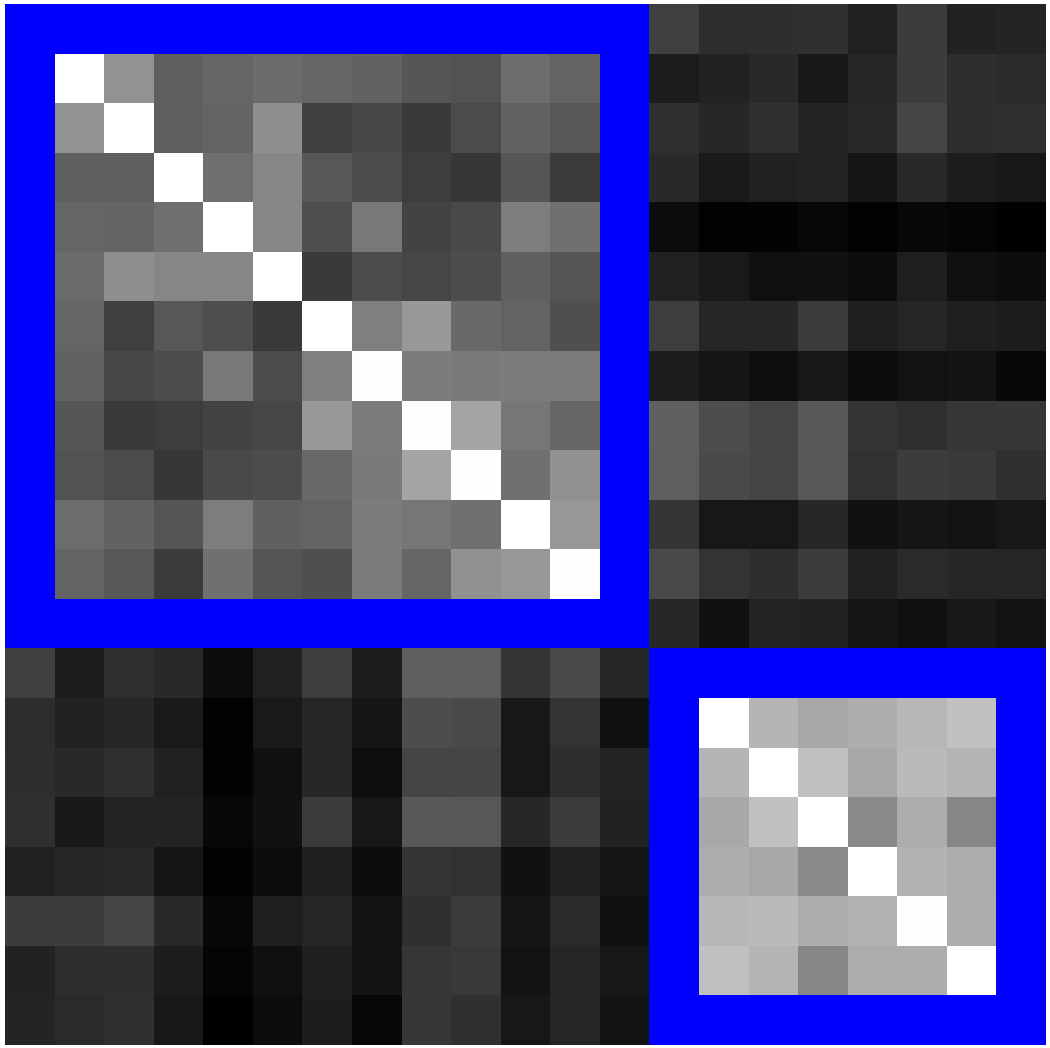

Clustering of OA and RA (2 clusters)
Principle Comp 3
30
20
10
0
-10
-20
20
0
-20
-40
Principle Comp 2
-40
-20
0
20
40
Principle Comp 1


Clustering of OA and RA (3 clusters)
Principle Comp 3
30
20
10
0
-10
-20
20
0
-20
-40
Principle Comp 2
-40
-20
0
20
40
Principle Comp 1


Clustering of OA and RA (4 clusters)
Principle Comp 3
30
20
10
0
-10
-20
20
0
-20
-40
Principle Comp 2
-40
-20
0
20
40
Principle Comp 1


Clustering of OA and RA (5 clusters)
Principle Comp 3
30
20
10
0
-10
-20
20
0
-20
-40
Principle Comp 2
-40
-20
0
20
40
Principle Comp 1

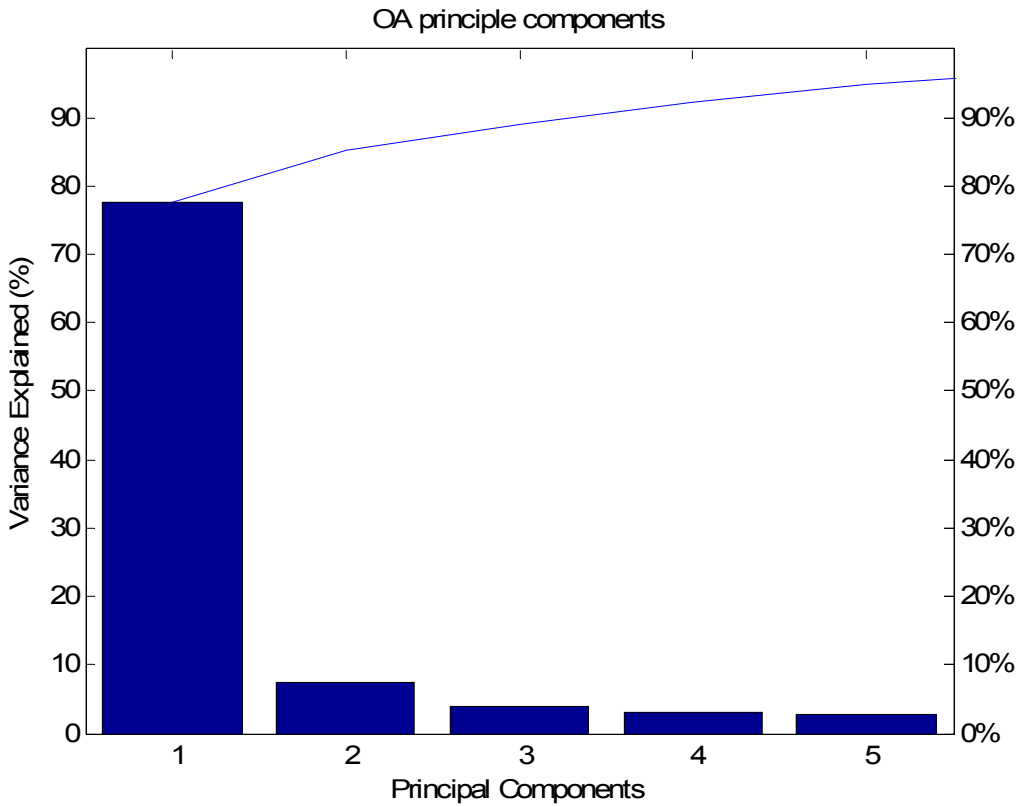
Spectral Clustering

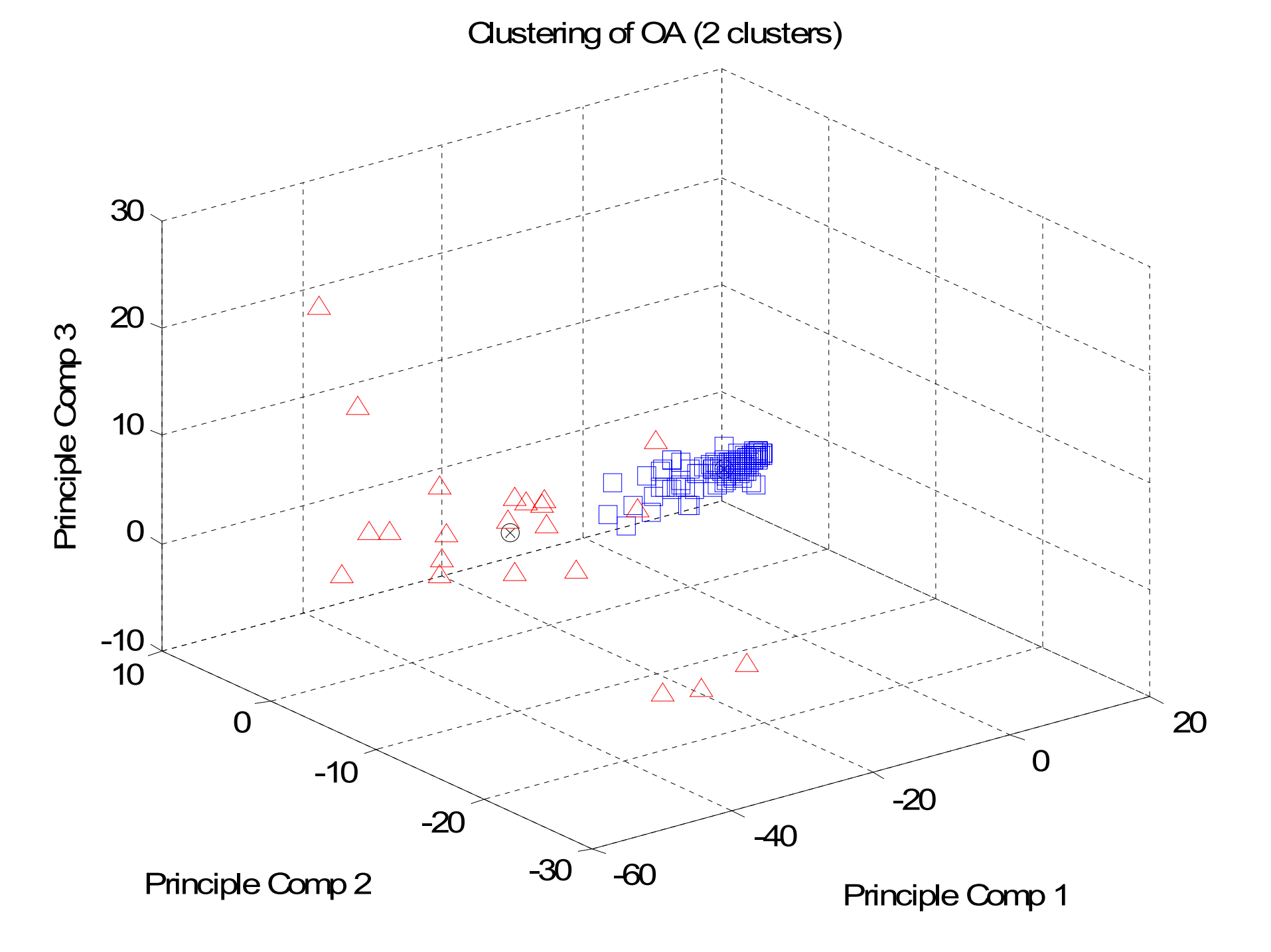


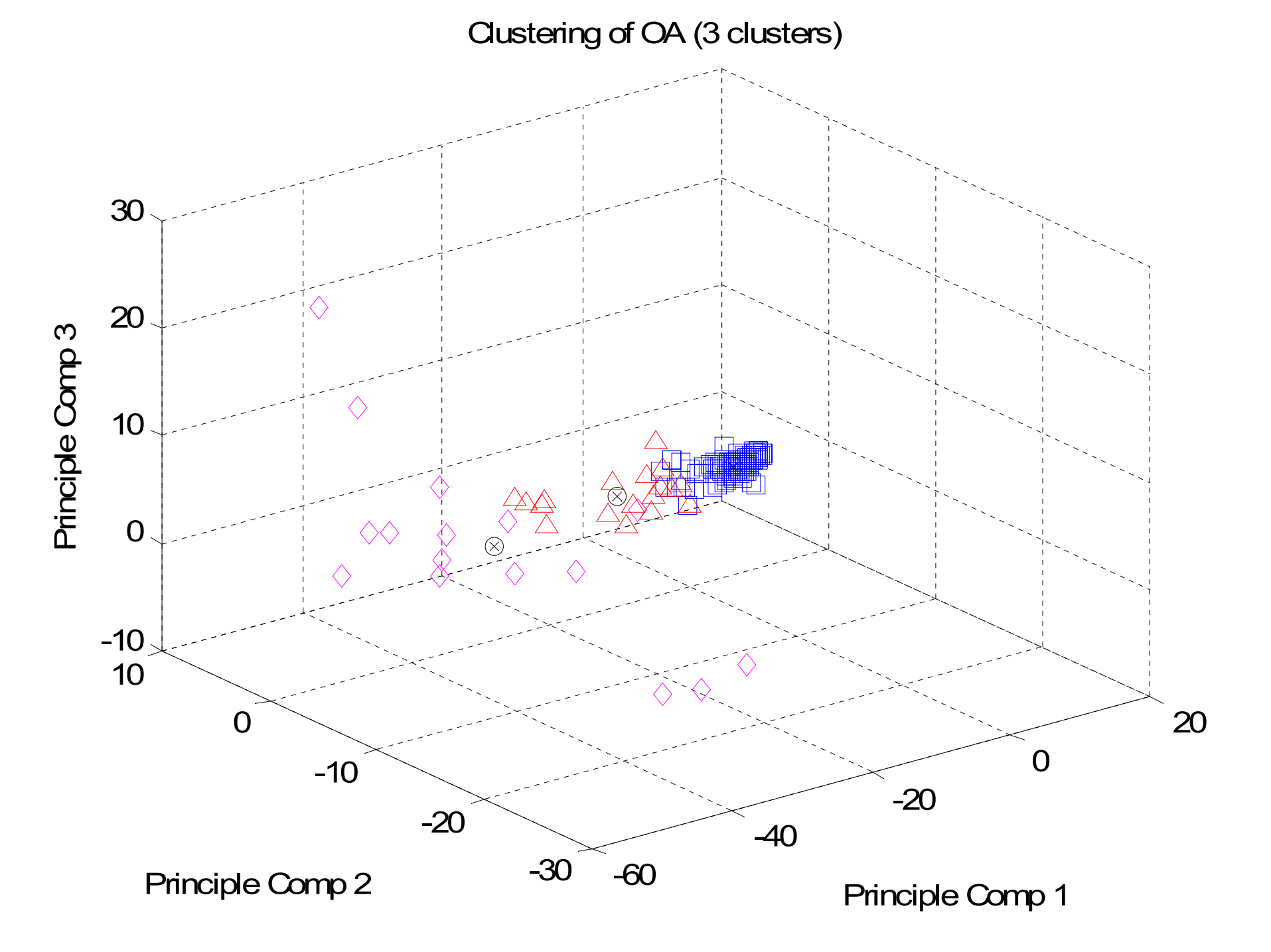
Clusters

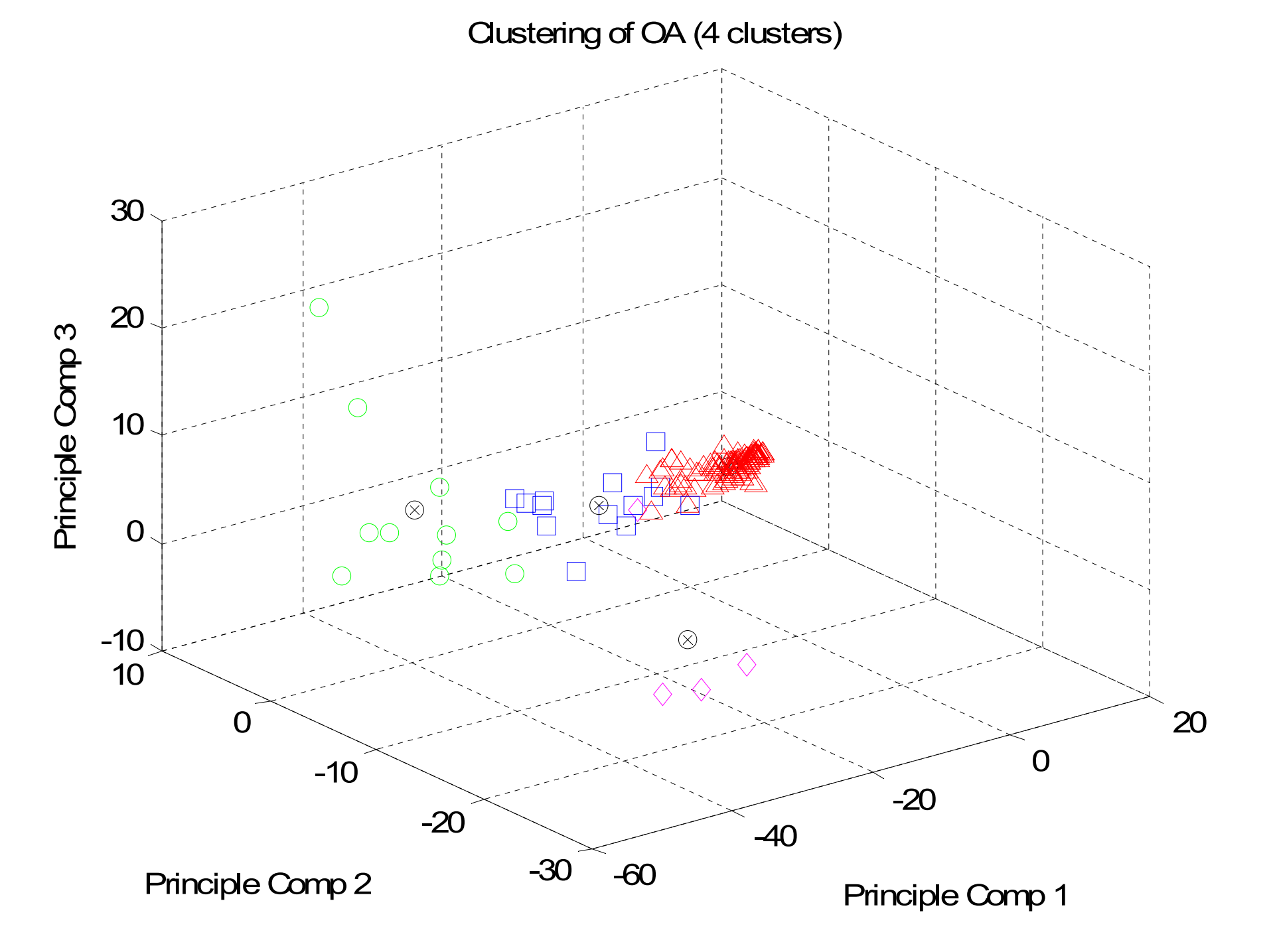
- Spectral
  - Cluster 1
    - 13 OA patients
  - Cluster 2
    - 8 RA patients
- K-means
  - 2 Clusters (13 OA patients, 8 RA patients)
  - 3 Clusters (9 OA, 8 RA, 4 OA)
  - 4 Clusters (7 OA, 8 RA, 4 OA, 2 OA)

Clustering Genes (OA)

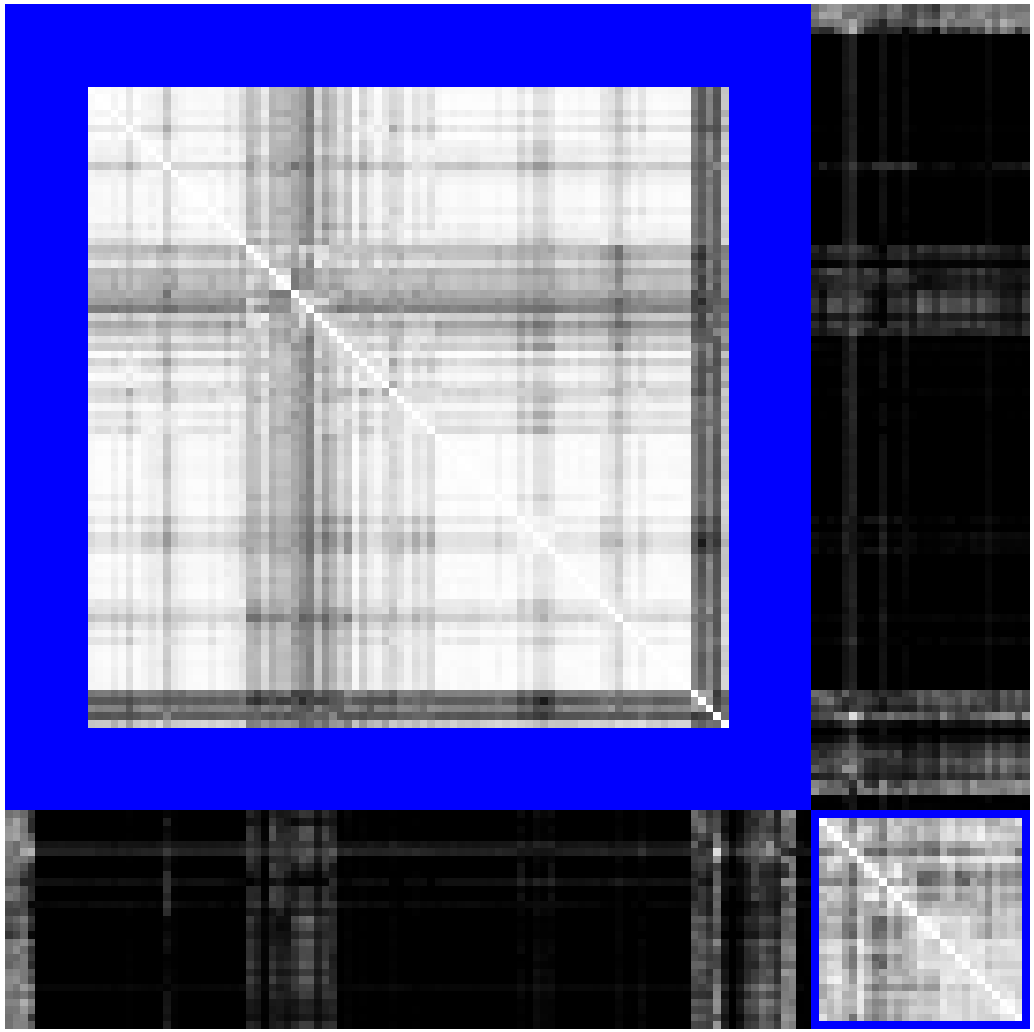



Clustering of OA (2 clusters)
Principle Comp 3
30
20
10
0
-10
10
0
-10
-20
-30
Principle Comp 2
-60
-40
-20
0
20
Principle Comp 1


Clustering of OA (3 clusters)
Principle Comp 3
30
20
10
0
-10
10
0
-10
-20
-30
Principle Comp 2
-60
-40
-20
0
20
Principle Comp 1


Clustering of OA (4 clusters)
Principle Comp 3
30
20
10
0
-10
10
0
-10
-20
-30
Principle Comp 2
20
0
-20
-40
-60
Principle Comp 1

Spectral Clustering



Clusters

- Spectral Clustering
  - Genes in Cluster 1
    - RNB6, FLJ10342, CDK7, BRD3
  - Genes in Cluster 2
    - CTSD, TIMP2, FUBP1
- K-means (2 Clusters)
  - Genes in Cluster 1
    - TIMP2, CDK7
  - Genes in Cluster 2
    - CTSD, RNB6, FLJ10342

?