Robust Clustering of PET data

Prasanna K Velamuru

Dr. Hongbin Guo (Postdoctoral Researcher),

Dr. Rosemary Renaut (Director, Computational Biosciences Program,)

Dr. Kewei Chen (Banner Health PET Center, AZ)

Dept. of Mathematics & Statistics

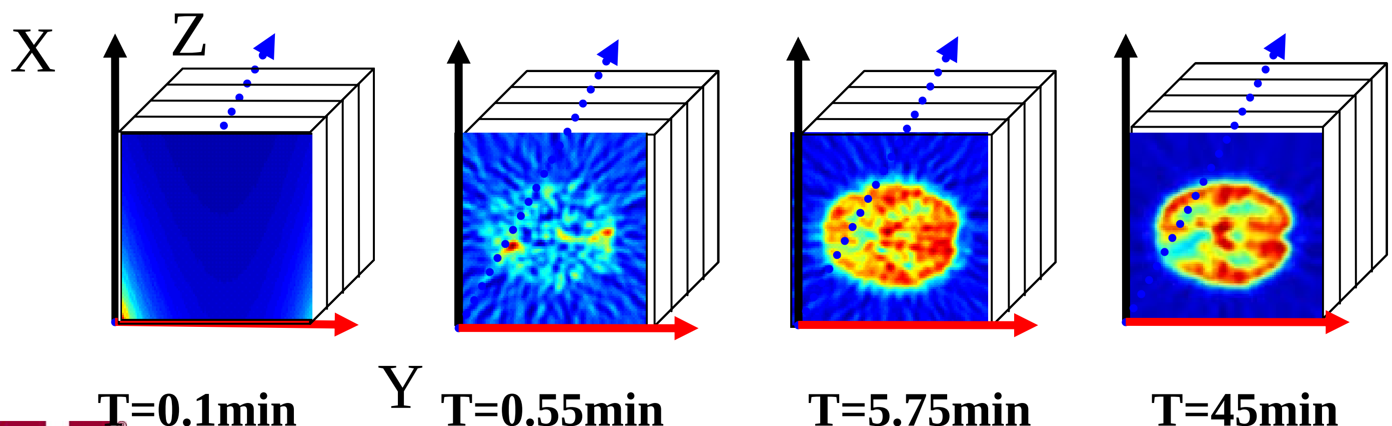
Arizona State University, Tempe, AZ

Outline

- Overview of data used and relevant terms
- Rationale behind clustering PET data
- Clustering Integrals/TACs
- Clustering Algorithms used
- Validation Measures used
- Results
- Conclusions

Dynamic Human Brain PET data

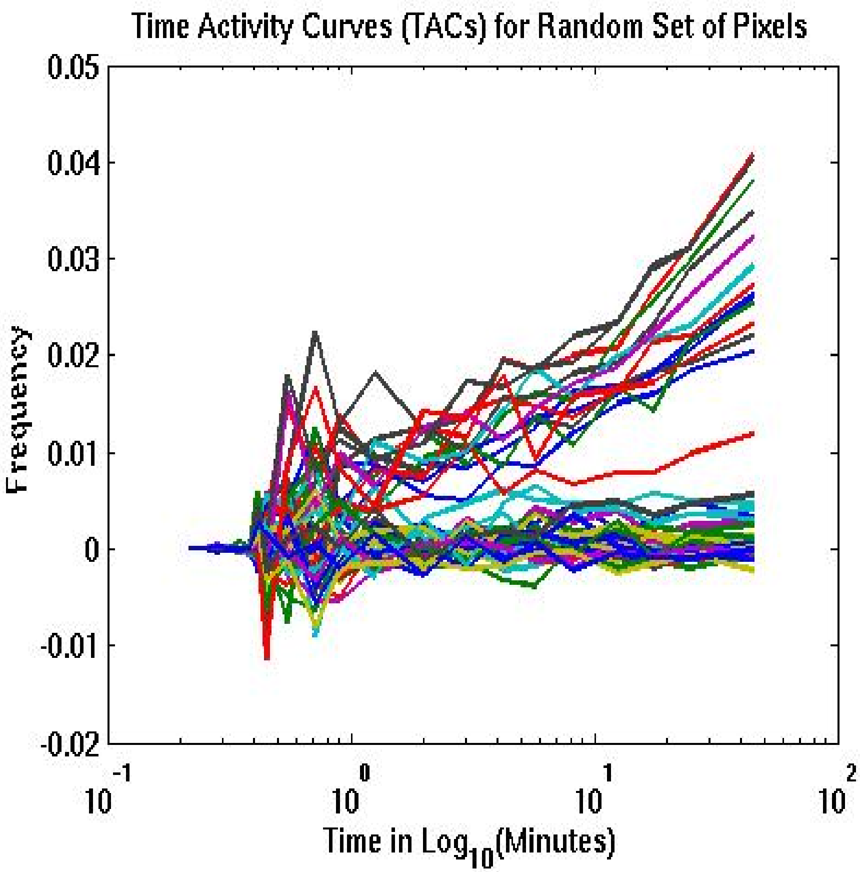
- Scanner: Siemens 951/31
- Output Image Matrix: 128 x 128 x 31 x 21
- 4-Dimensional: 3 in Space, 1 in Time.



Time Activity Curves (TACs)

- Vector: $[X_1, X_2, \ldots, X_{20}, X_{21}]$
- Defined for each Voxel
- Total for each slice: 128*128=16384
- For Entire brain Volume:

  128*128*31= 507904



Integrals

- Product of individual TACs with scanning time durations. Biologically relevant – accumulation of tracer in tissue over time.

$[X_1, X_2, \ldots, X_{20}, X_{21}]$ *

8 * 0.0333
2 * 0.1667
0.2000
0.5000
2 *1.0000
2 *1.5000
3.5000
2* 5.0000
10.0000
30.0000

- Single **Scalar** value
- Approximate Estimate
- Reduces Dimension of data

Clustering PET data- Why is it important?

- Output data is very noisy
  - Low Signal to Noise Ratio (SNR)
- Important preprocessing step performed prior to parametric estimation from dynamic PET data.
- Form of Segmentation
- Provide better information
  - Partially overcome the effect of noise in the data
  - Improve accuracy of voxel level quantification of parametric data from PET images

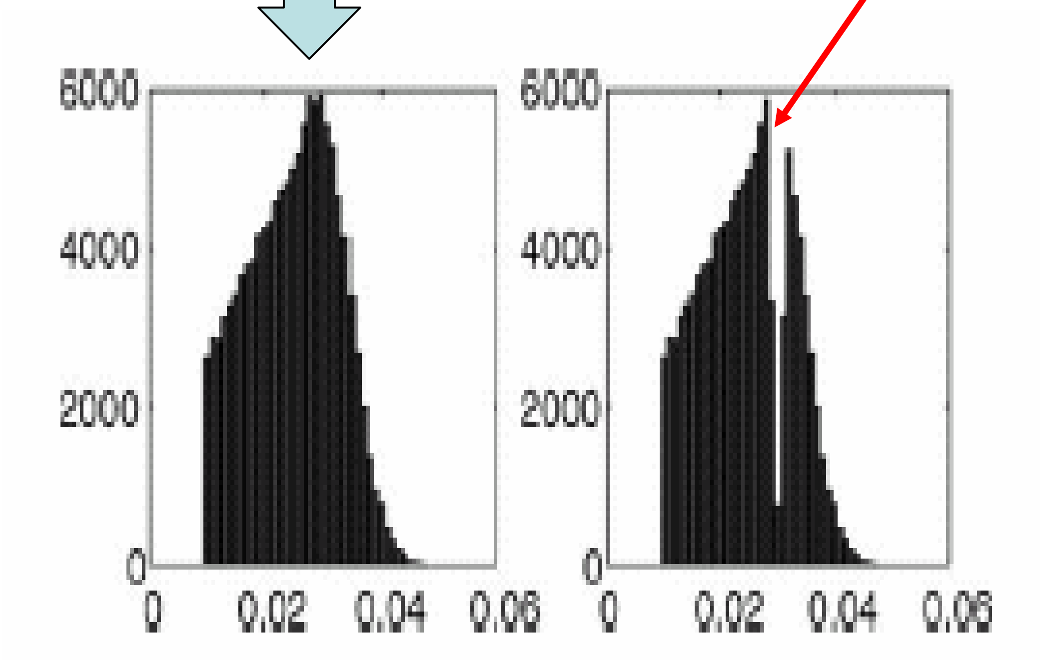
Clustering Algorithms
(Jain and Dubes, 1988)

- Hierarchical
  - Average Linkage (HAL)
  - Centroid Linkage (HCL)
  - HAL & HCL with preclustering (HAL1/HCL1)
  - HAL & HCL with preclustering and assignment of isolated voxels (HAL2/HCL2)
- Partitional
  - K-Means

Preclustering: Illustration

Frequency=5934
Density~0.03

Precluster



Clustering Algorithms for PET: Intangibles

- Unsupervised
  - no predefined classes
  - no specific examples that would show what kind of important relationships within the data are of biological significance.
- Optimal Number of clusters
  - Not known *a priori* (very difficult to set up simulated data)
  - Difficult to rely on visual perception due to high noise and dimensionality
- How appropriate is the clustering method for the data at hand?
- Trade-off between computational cost and cluster quality
  - How much tolerance is accepted?

Validation Measures Used

- Computed at the element, cluster and global level for each algorithm.

Intra Cluster Measures	Inter Cluster Measures
•Average Distance to Mean •Maximum Dist. to Mean •Maximum Diameter •Average Spread •Total Energy	•Separation •Minimum Separation •Average Split

Validation Measures Used

- Average Silhouette Width

$$c_i^j = \frac{b_i^j - a_i^j}{max(a_i^j, b_i^j)}$$

- Dunn's Index

$$v_{GD}(U) = \underbrace{\min_{1 \leq i \leq c}} \left\{ \underbrace{\min_{1 \leq j_{j \neq i} \leq c}} \left\{ \frac{\delta(X_i, X_j)}{\max_{\underbrace{1 \leq k \leq c}} \{\Delta(X_k)\}} \right\} \right\}$$

- Average Linkage (Bezdek et.al. ,1998, δ_3)

$$\delta_{AL}(X_j, X_k) = \frac{1}{n_j n_k} \sum_{x \in X_j, y \in X_k} d(x, y)$$

- Complete Linkage (Bezdek et.al.,1998, δ_2)

$$\delta_{CL}(X_j, X_k) = \max_{x \in X_j, y \in X_k} \{d(x, y)\}$$

- Combined Average Linkage (Bezdek et.al.,1998, δ_5)

$$\delta_{CAL}(X_j, X_k) = \frac{1}{n_j + n_k} \{ \sum_{x \in X_j} d(x, \mu_k) + \sum_{y \in X_k} d(y, \mu_j) \}$$

Average Spread

Very Less Difference observed

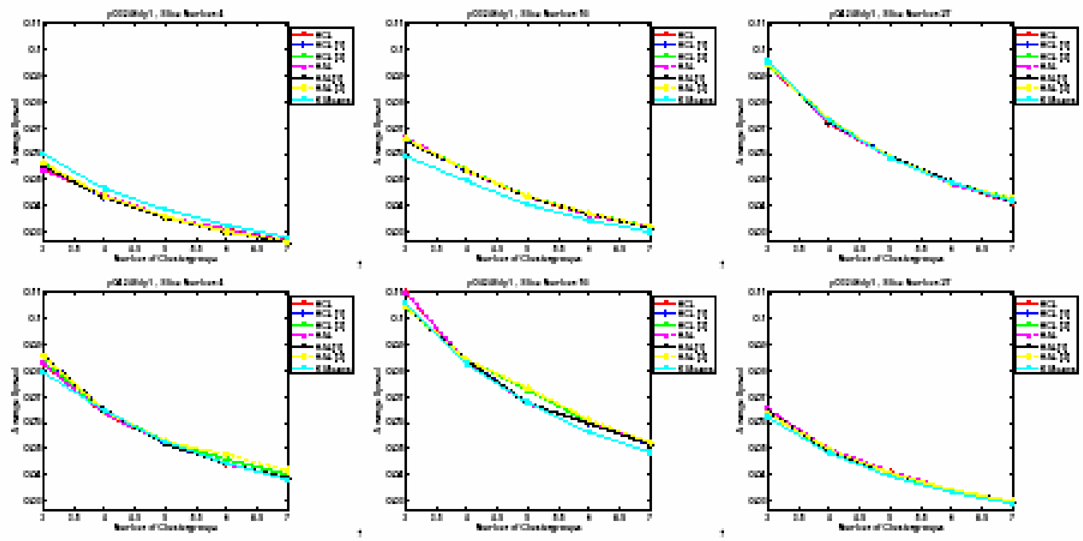


Figure 14. Average Spread Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (Integrals)

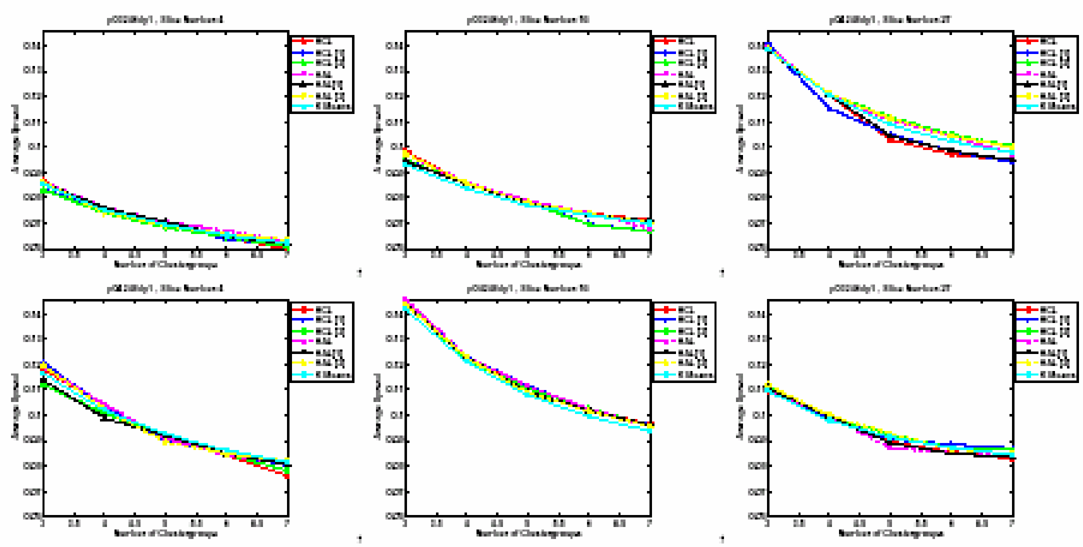


Figure 15. Average Spread Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (TACs)

Average Silhouette Width

Observe maxima across each slice for each subject for optimal groups

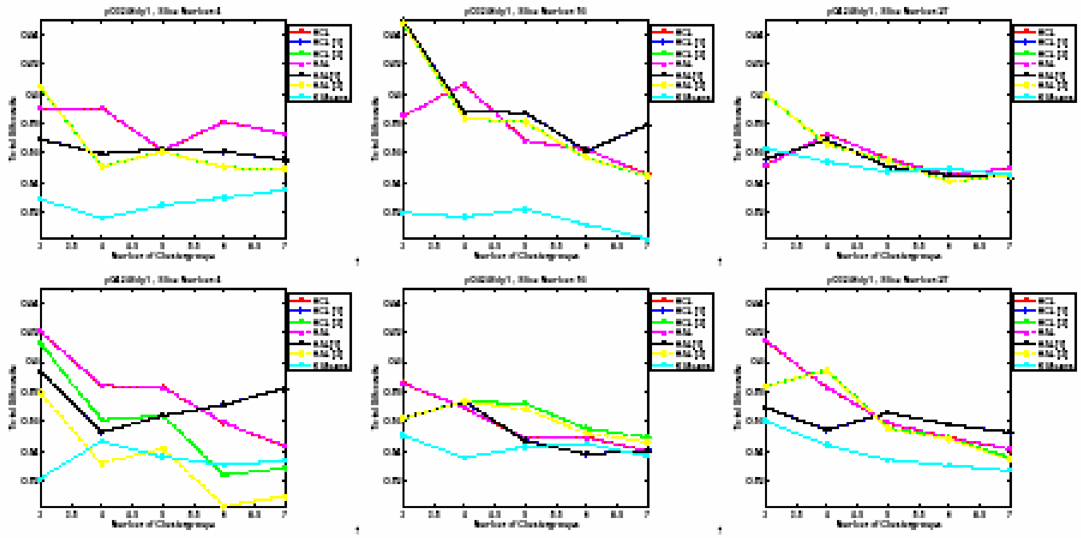


Figure 24. Silhouette Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (Integrals)

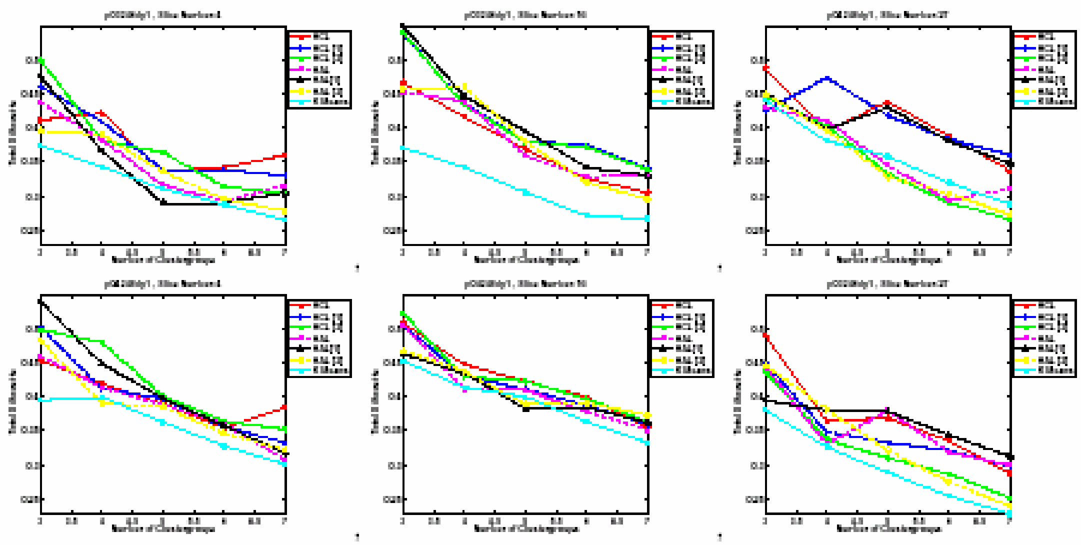


Figure 25. Silhouette Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (TACs)

Combined Average Ratio

Integrals- More pronounced maxima compared to TACs

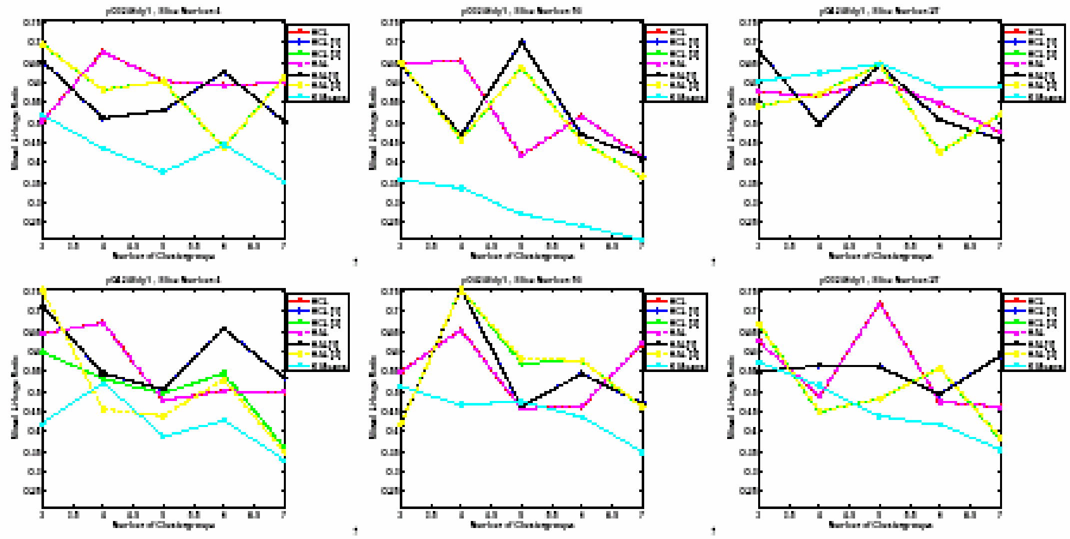


Figure 30. Combined Average Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (Integrals)

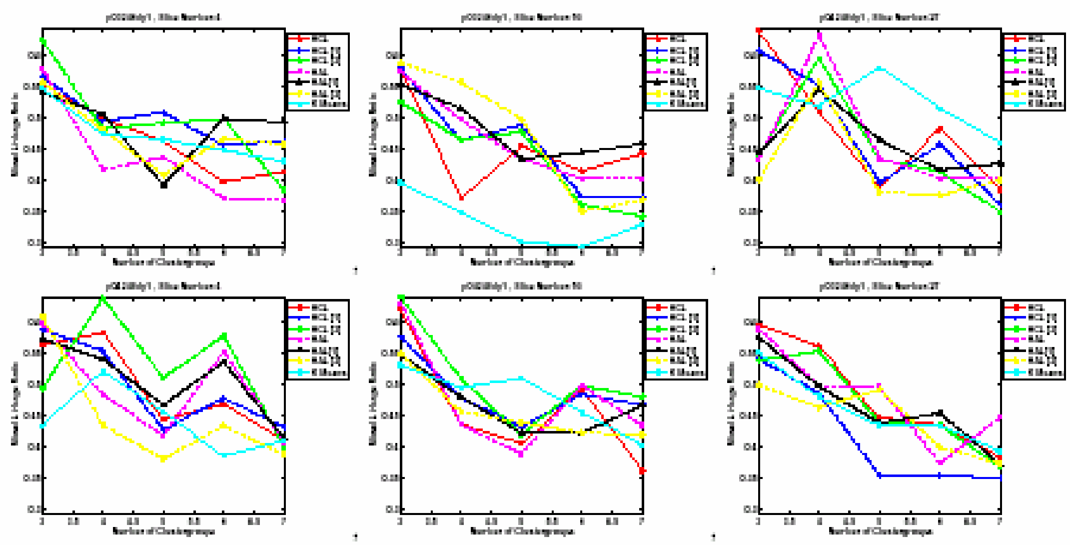


Figure 31. Combined Average Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248 (TACs)

Conclusions

- TAC vs Integral Data
  - Significant difference in intra and inter cluster measures ***NOT*** observed
  - Inter and Intra cluster measures ***relatively higher*** for TAC data
  - Difference in Intra-cluster measures **>** Difference in Inter-cluster measures
  - If more well separated clusters desired – need to use multidimensional TAC data

Conclusions

- Fast Hierarchical methods ***comparable*** to more expensive standard forms
- K-means
  - Greater average dissimilarity within clusters
- Inter-Intra Cluster Measures
  - Integral data : more pronounced maxima
- Setting number of clusters *a-priori*
  - Not appropriate
  - Optimal number of clusters varies depending upon slice and subject

Future Directions

- Use results of clustering to find parametric values
- Assess cluster validity through parametric values
- Use cluster ranges on parametric values as constraints for voxel wide parameter estimations by nonlinear optimization

References

- Phelps, M.E., Positron Emission Tomography. In: Mazziotta, J. and Gilman, S., Eds.,1992, ***Clinical Brain Imaging: Principles and Applications***
- Hongbin Guo, Rosemary Renaut, Kewei Chen and Eric Reiman,2003, ***Clustering Huge Data sets for Parametric PET imaging***, Biosystems, 71, 1-2, pp.81-92
- A.K. Jain and R. C. Dubes. ***Algorithms for Clustering Data***. Prentice Hall, 1988.
- Jain, A.K., Murty M.N., and Flynn P.J. (1999): ***Clustering: A Review***, ACM Computing Surveys, Vol 31, No. 3, 264-323
- Kaufman L., and Rousseeuw P., 1990. Finding groups in Data : ***An introduction to Cluster Analysis,*** John Wiley and Sons, New York, NY.
- Everitt B.S., Landau S., Leese M., 2001. ***Cluster Analysis***, 4 th Edition. Edward Arnold, London, UK.
- Halkidi, M., Batistakis, Y., Vazirgiannis M., 2001. ***On Clustering Validation Techniques,*** Journal of Intelligent Information Systems, 17:2/3, 107–145.
- Kimura, Y., Senda, M., Alpert, N.,2002, ***Fast formation of statistically reliable FDG parametric images based on clustering and principal components***, Phys. Med. Biol. 47(3),pp.455-468
- Velamuru, P.K., ***Validation of clustering algorithms for clustering PET data***, CBS Internship Report, http://lifesciences.asu.edu/compbiosci/stext/students.htm#projects

Acknowledgements

- Dr. Rosemary A Renaut

  Director, Computational Biosciences Program, Dept. of Mathematics & Statistics, ASU, Tempe, AZ
- Dr. Hongbin Guo

  Post-doctoral Researcher, Dept. of Mathematics & Statistics, ASU, Tempe, AZ
- Dr.Kewei Chen

  Bio-mathematician, Banner Good Samaritan Hospital PET Center, Phoenix, AZ