
Validation of Clustering Methods for PET data

Prasanna K Velamuru

Advisors:

Dr. Rosemary A Renaut

Dr. Hongbin Guo

Dr Huan Liu

Outline

- Motivation
 - Clustering Algorithms Validated
 - Validation Metrics
 - Results
 - Significance of Results
 - Conclusions and Future Work
-

PET – Quick Overview

- Diagnostic Imaging technique
 - Functional Information
 - Knowledge of biochemical basis of both normal as well as abnormal functions
 - Integral part of clinical care
 - Oncology, Cardiology, Neurology/Psychiatry
-

Domain Of PET data used in this Study

- Measurement of biochemical and physiological parameters from dynamic human brain PET data.
- Three-compartment FDG tracer kinetic model

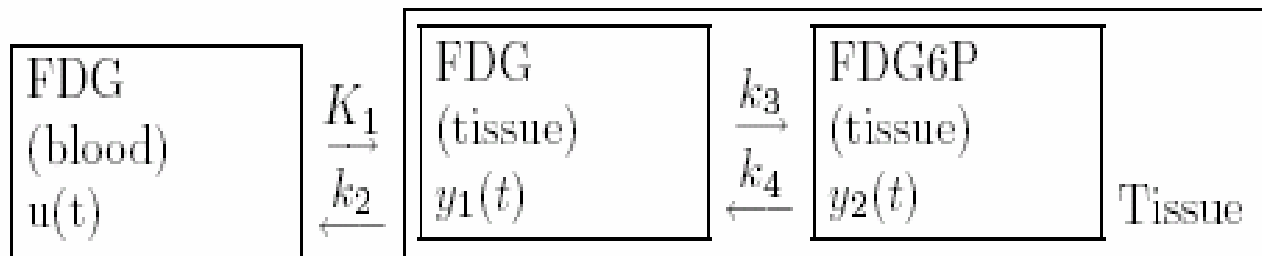
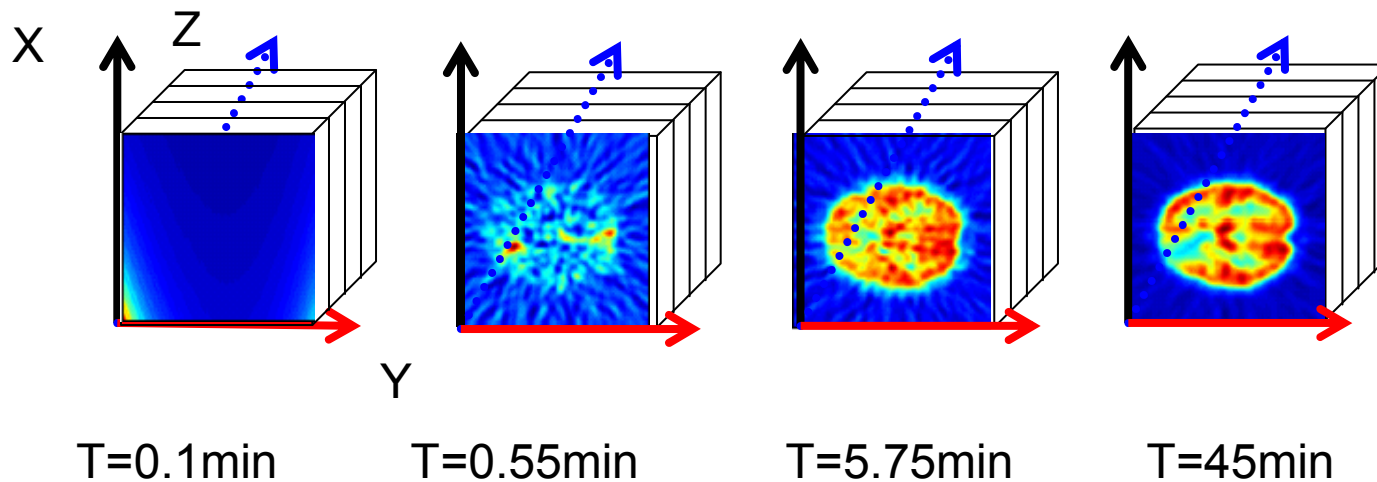


Figure 1. FDG tracer kinetic model

Dynamic Human Brain PET Data

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

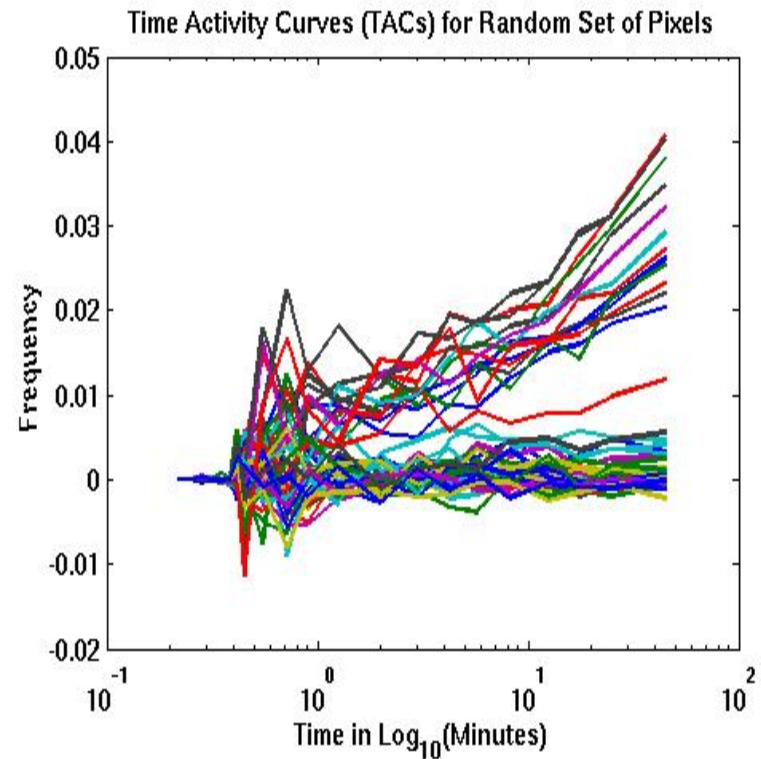


Characteristics of Dynamic PET data

- Output data is very noisy
 - Low Signal to Noise Ratio (SNR)
 - Data obtained at irregular time intervals
 - Initial time intervals
 - Very short: to observe gradient in the output
 - last time interval
 - Long
 - Output changes little
 - Useful for determining long term decay term in the output
 - Contributes significantly in the estimation of the kinetic rates.
-

Time Activity Curves

- Vector: $[X_1, X_2, \dots, X_{20}, X_{21}]$
- Defined for each Voxel / Pixel
- Total for each slice:
 $128 * 128 = 16384$ TACs
- For Entire brain Volume:
 $128 * 128 * 31 = 507904$



Integrals

- Obtained by multiplying individual TACs with time intervals matrix.

$$[X_1, X_2, \dots, X_{20}, X_{21}] * \begin{pmatrix} 0.0333 \\ 0.0333 \\ \vdots \\ \vdots \\ 0.5000 \\ \vdots \\ 1.5000 \\ \vdots \\ 10.000 \\ 30.000 \end{pmatrix}$$

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

Clustering: In context of PET data

- Important preprocessing step performed prior to parametric estimation from dynamic PET data.

Original Image, Subject:3246, Slice16

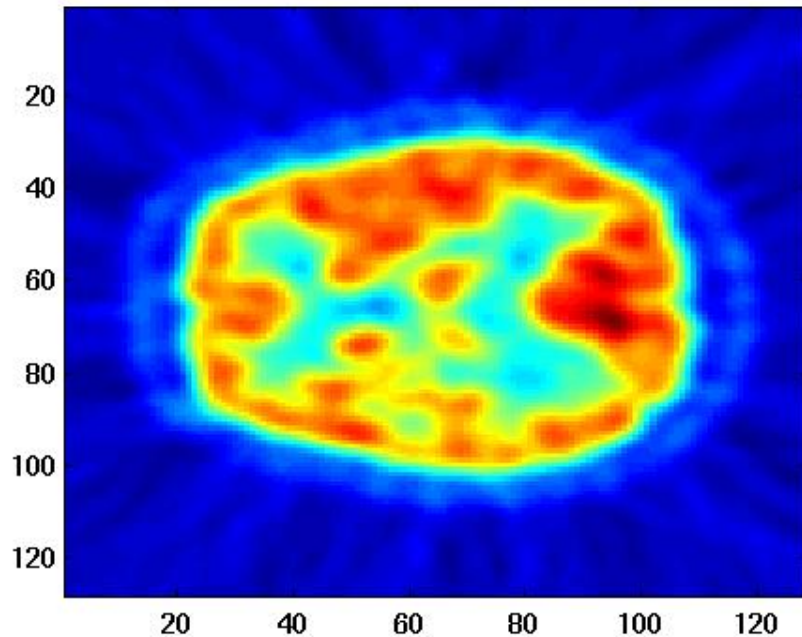
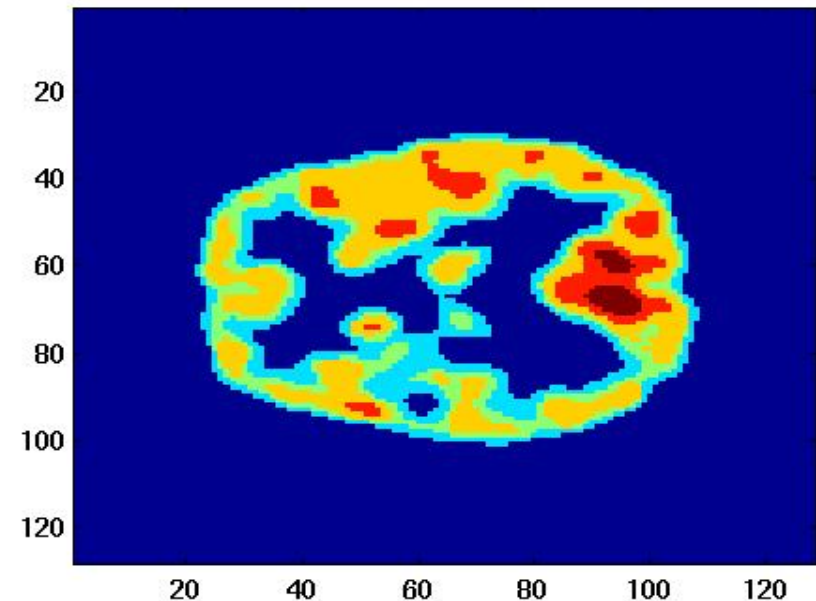


Image showing 5 clusters, Method: HAL, Subject:3246, Slice16



Motivation

- New preprocessing clustering techniques designed and published by Guo et al
 - Significantly reduces the overall time for clustering
 - Basic initial validation done in the past
 - Not comprehensively compared and validated with classical clustering methods
-

Preliminary Concepts: Distance Measures

- Measurement of (dis)similarity between multivariable vectors / integrals
 - Weighted Minkowski p norm
$$\left(\sum_{l=1}^n |(x_l - y_l)w_l|^p \right)^{1/p}$$
 - Distances weighted by the time duration intervals $([\Delta t_1, \Delta t_2, \dots, \Delta t_{20}, \Delta t_{21}])$
 - Weighted Euclidean (“ L_2 ”) and Weighted Manhattan (“ L_1 ”)
 - Found to be good distance measures for PET
-

Preliminary Concepts: Histogram-based Thresholding

Properties of Last Frame

- Provides more accurate image reconstruction
- Higher SNR compared to short time interval frames
- FDG accumulation is higher
- Distribution of voxel intensities
 - Provides closer correlation with spatial variation of trace in tissue.
- Supports thresholding to identify active voxels

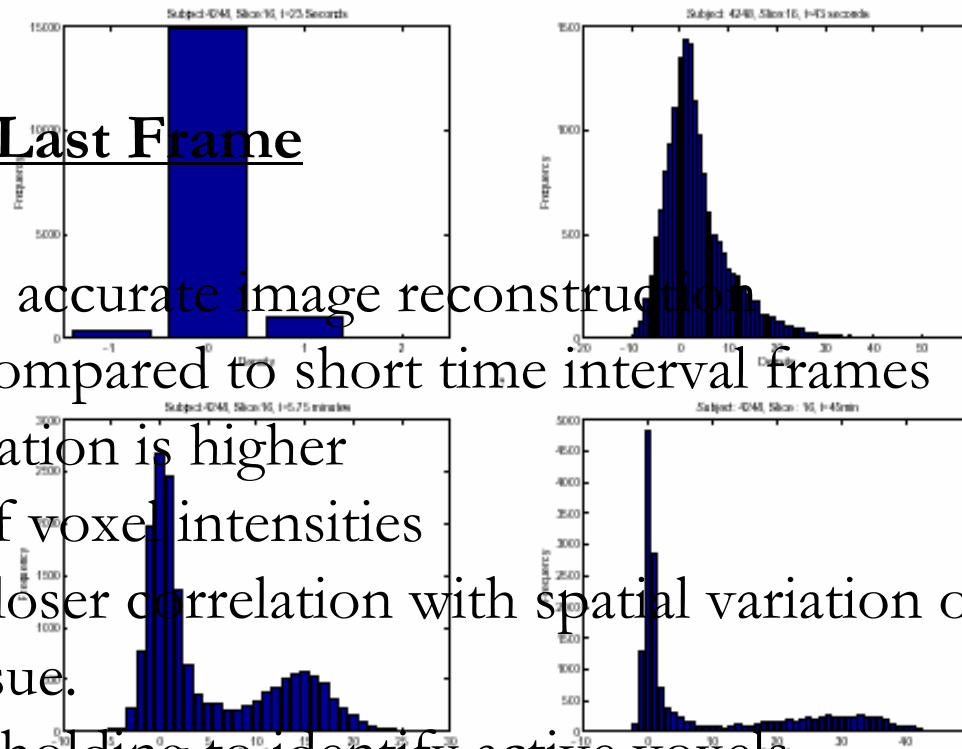


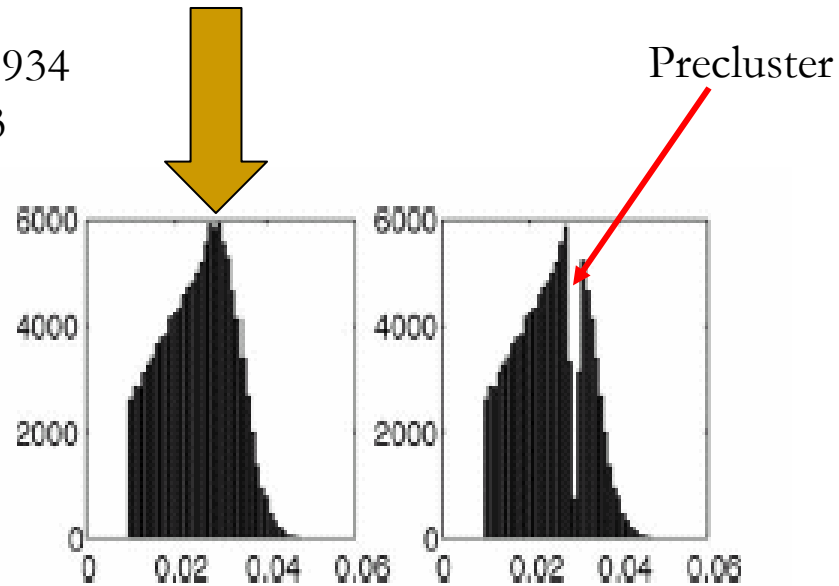
Figure 6. Histograms of densities summed over all three spatial dimensions for time frames with $t=$ 23 seconds, 43 seconds, 5.75 minutes and the last frame at time 45 min.

Preliminary Concepts: Preclustering

- Initial thresholding
 - Does not provide enough information to cluster data
 - Should include features over the entire set of TACs
 - Compute mean TAC
 - From set of TACs that contain voxels with highest frequency
 - Mean TAC used to find initial cluster
 - Search is performed over all voxels
 - Distances are measured with respect to the entire TAC
-

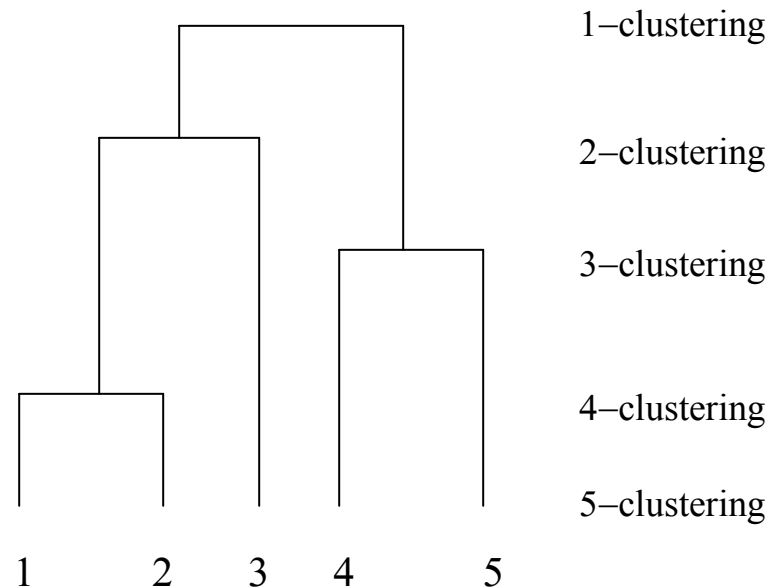
Preclustering: Illustration

Frequency=5934
Density~0.03



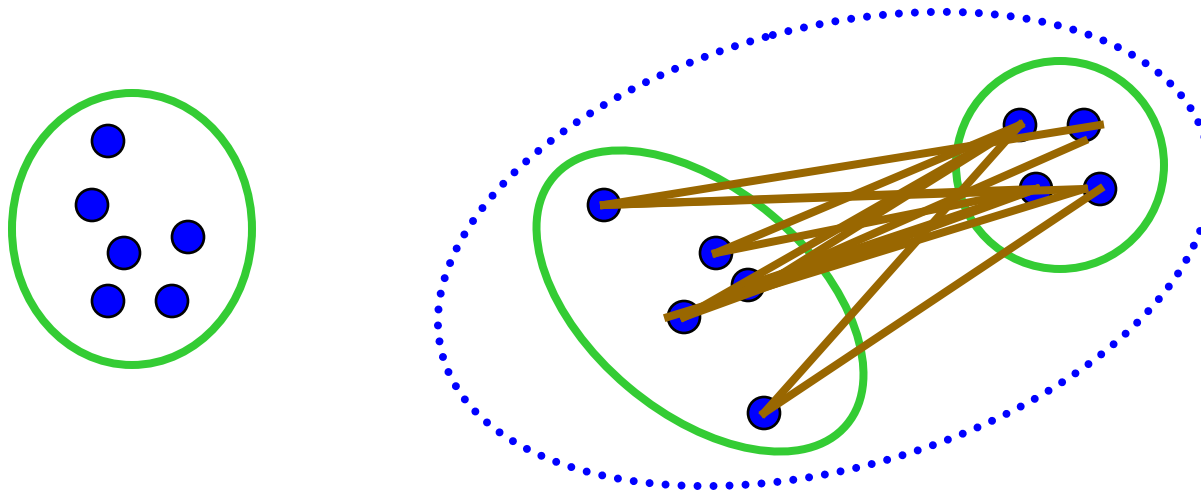
Clustering Algorithms Used for Validation: Hierarchical Clustering

- **Agglomerative** (bottom-up)
- **Algorithm:**
 - Initialize: each item as a cluster
 - Iterate:
 - select two *most similar* clusters
 - merge them
 - Halt: when required number of clusters is reached



Clustering Algorithms Used for Validation: HAL

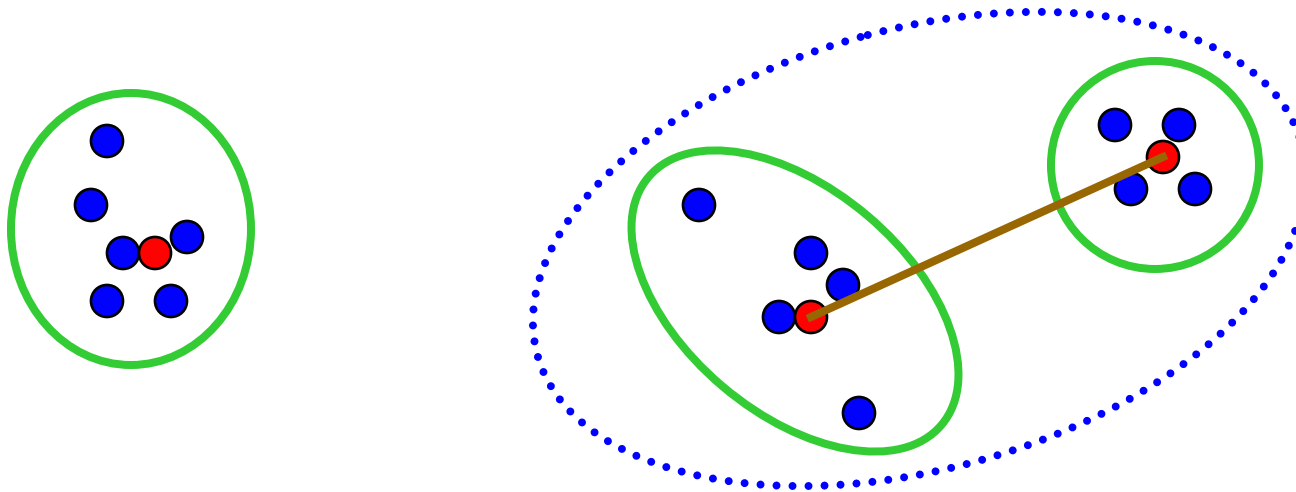
- Average Linkage
 - cluster similarity = **average** similarity of all pairs



$$d(C_1, C_2) = \frac{1}{n_{C_1} n_{C_2}} \sum_{i=1}^{n_{C_1}} \sum_{j=1}^{n_{C_2}} d(x_{C_1(i)}, x_{C_2(j)})$$

Clustering Algorithms Used for Validation: HCL

- Centroid Linkage
 - cluster similarity = similarity between **centroids**



$$d(C_1, C_2) = d(\bar{x}_{C_1}, \bar{x}_{C_2})$$

where $\bar{x}_{C_1} = \frac{1}{n_{C_1}} \sum_{i=1}^{n_{C_1}} x_{C_1(i)}$ and $\bar{x}_{C_2} = \frac{1}{n_{C_2}} \sum_{j=1}^{n_{C_2}} x_{C_2(j)}$

Clustering Algorithms Used for Validation: HCL1/HAL1 (with Preclustering)

- Algorithm Outline:
 - Initialization Phase
 - Filter out active and inactive voxels
 - Thresholding using final frame
 - Determine preclusters within each active voxels interval
 - Repeat preclustering process
 - No more active voxel intervals are available
 - Maximum iteration number is reached
 - Perform HAL/HCL on reduced data set containing active voxels
 - Preclusters and isolated voxels
-

Clustering Algorithms Used for Validation: HCL2/HAL2 (with Preclustering and Merging)

■ Algorithm Outline

- Perform same steps for initialization and preclustering as in HCL1/HAL1
 - Calculate mean TAC for all preclusters
 - For all isolated voxels
 - Calculate distance to the mean TACs of preclusters
 - Merge voxel with precluster of closest similarity (minimum distance)
 - Perform HAL/HCL on reduced data set containing active voxels
 - Only preclusters
-

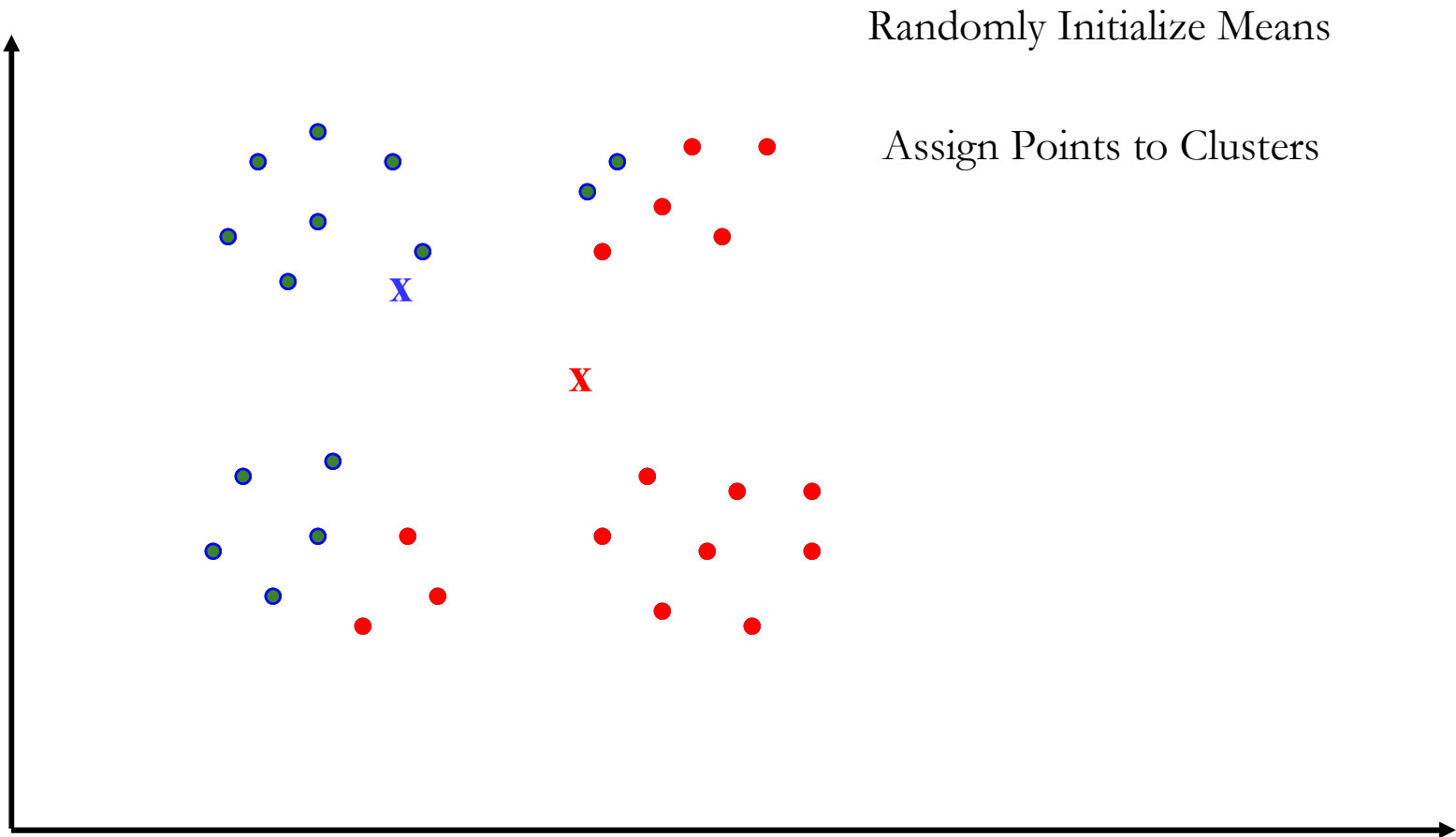
Clustering Algorithms Used for Validation: K-Means

- **Partitional** clustering algorithm
 - Iterative relocation
- Locally minimizes sum of squared distance (“energy”) between the data points and their corresponding cluster centers

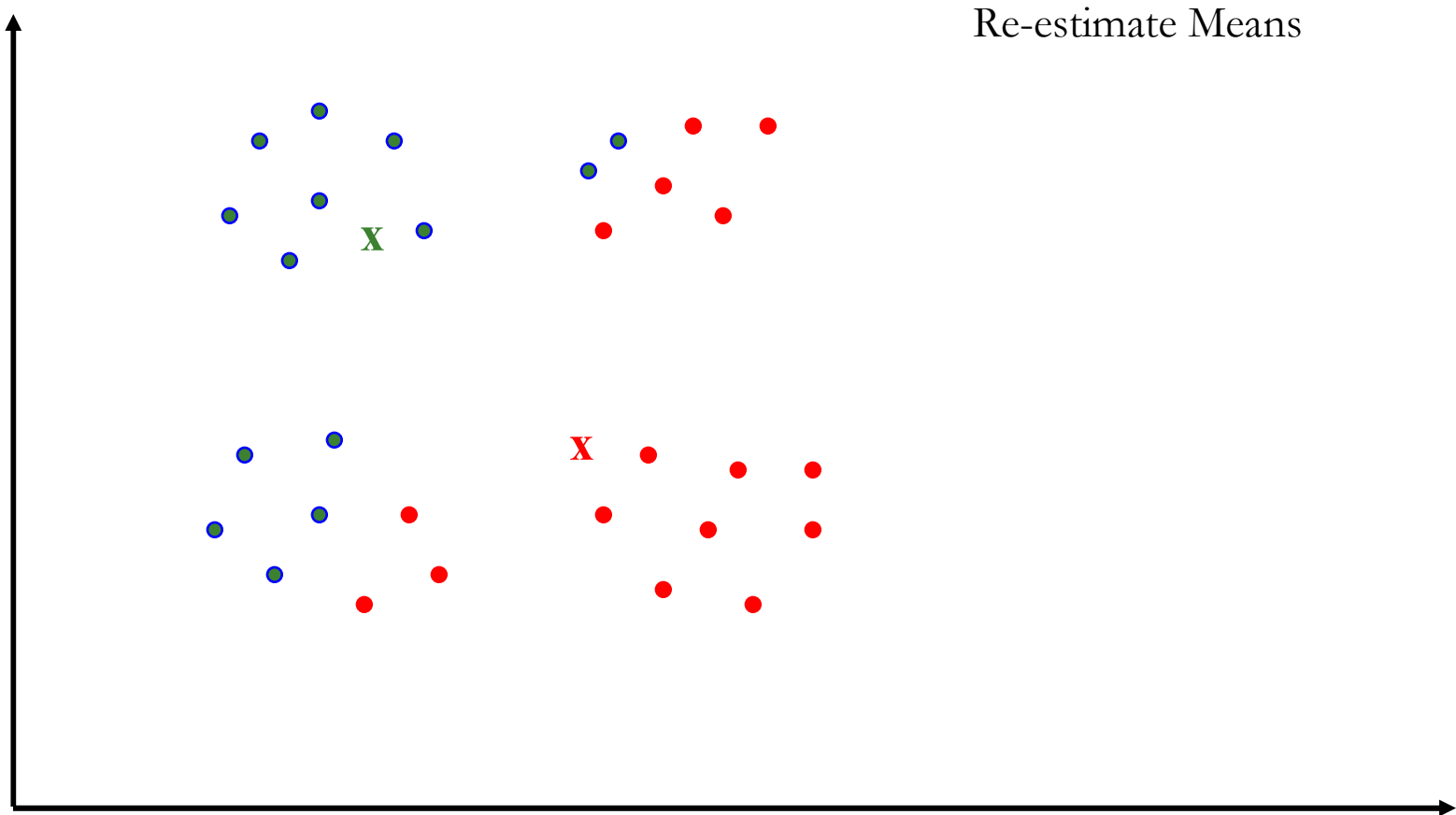
$$\sum_{l=1}^k \sum_{x_i \in X_l} [d(x_i, \mu_l)]^2$$

- Initialization of K cluster centers:
 - Totally random
 - Random perturbation from global mean
 - Heuristic to ensure well-separated centers
-

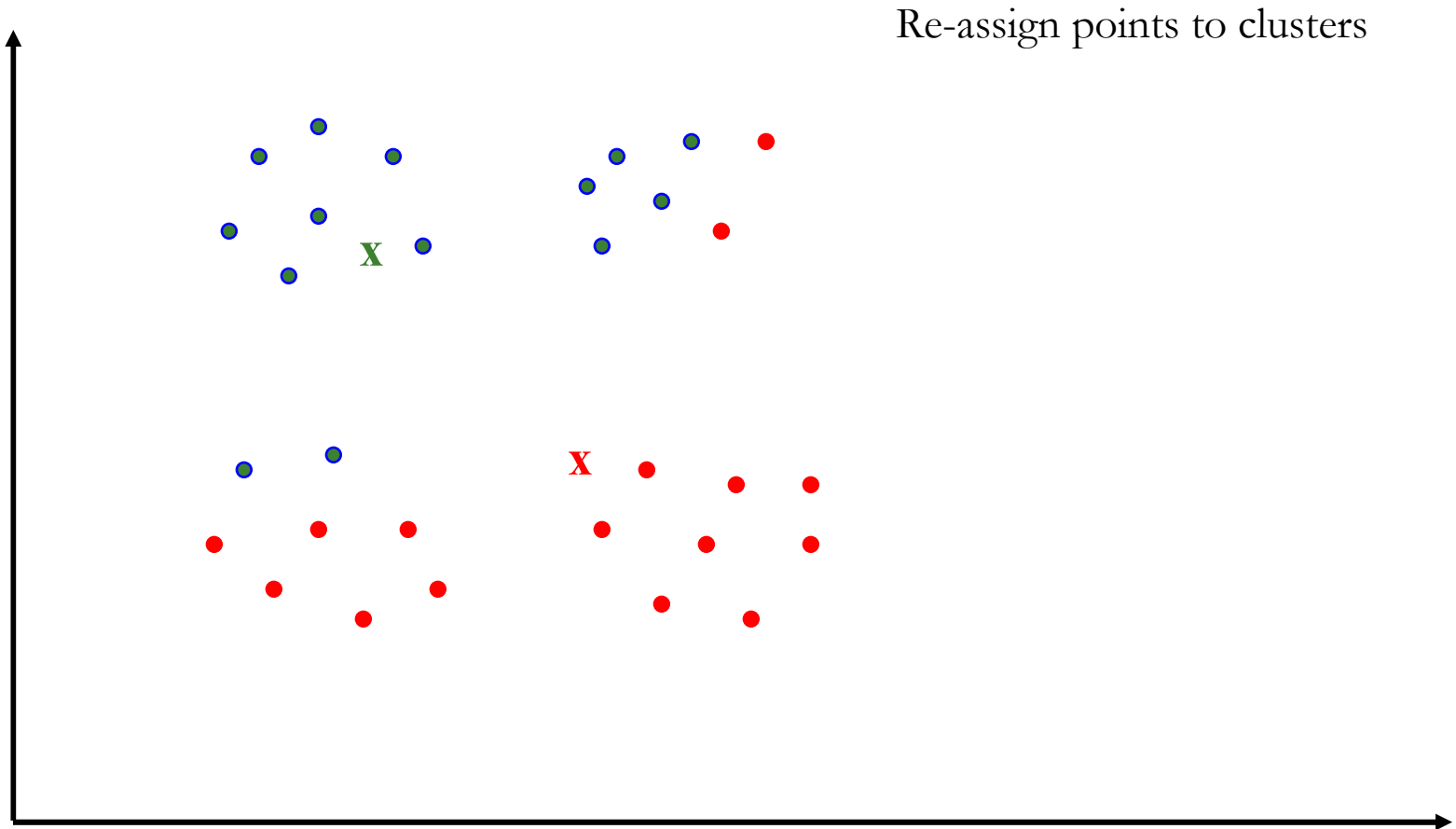
Clustering Algorithms Used for Validation: K-Means Illustration



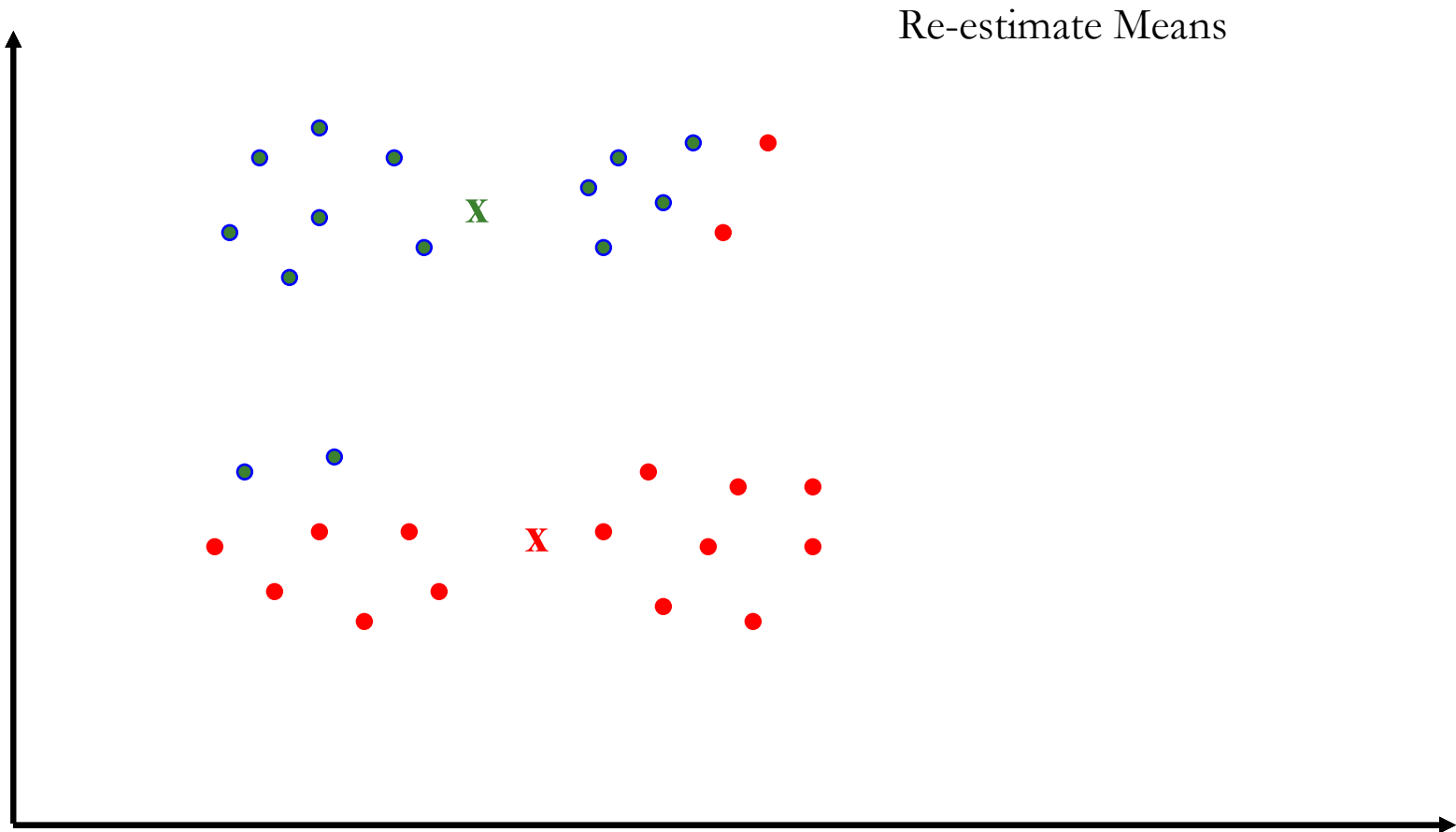
Clustering Algorithms Used for Validation: K-Means Illustration



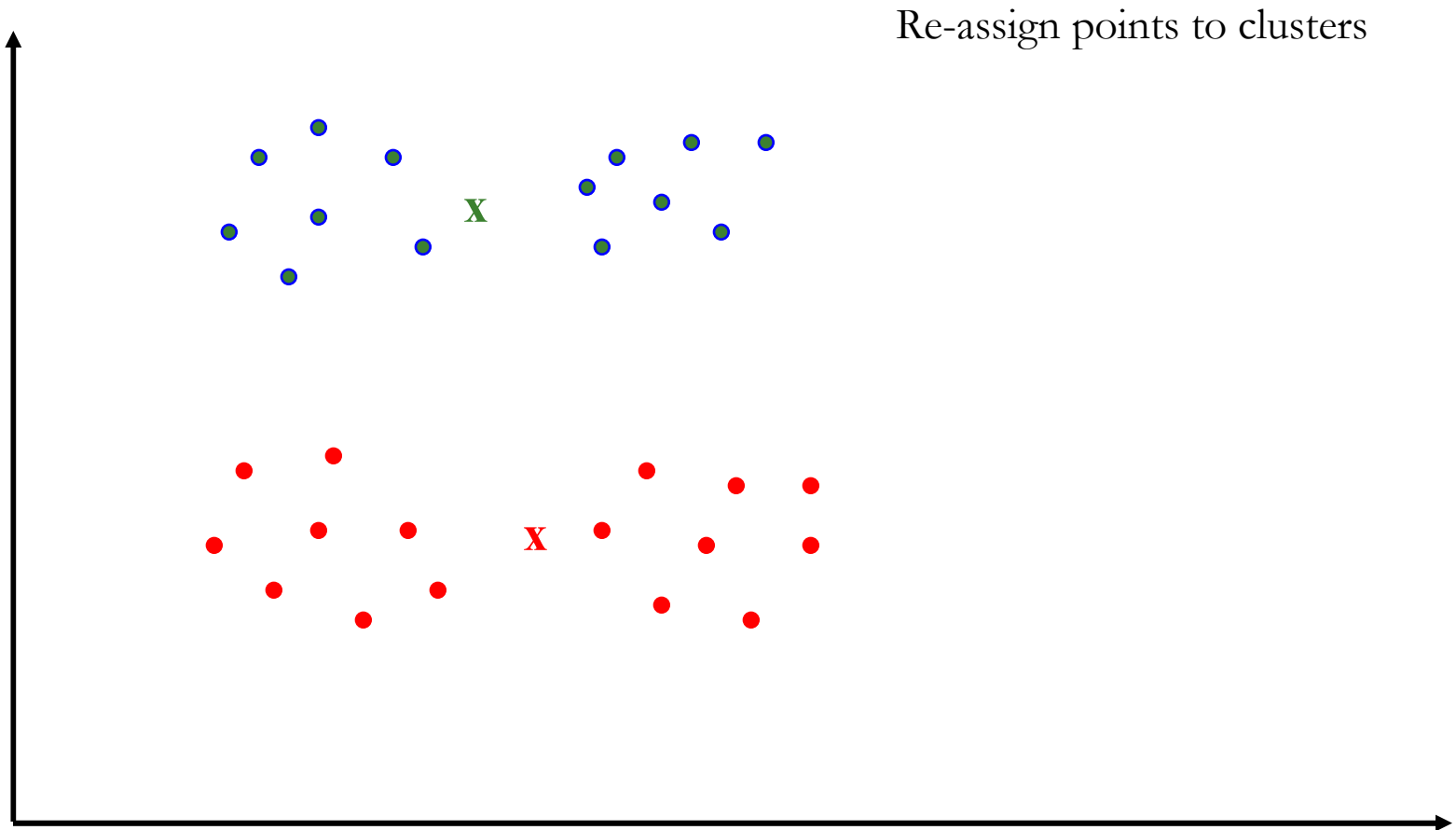
Clustering Algorithms Used for Validation: K-Means Illustration



Clustering Algorithms Used for Validation: K-Means Illustration

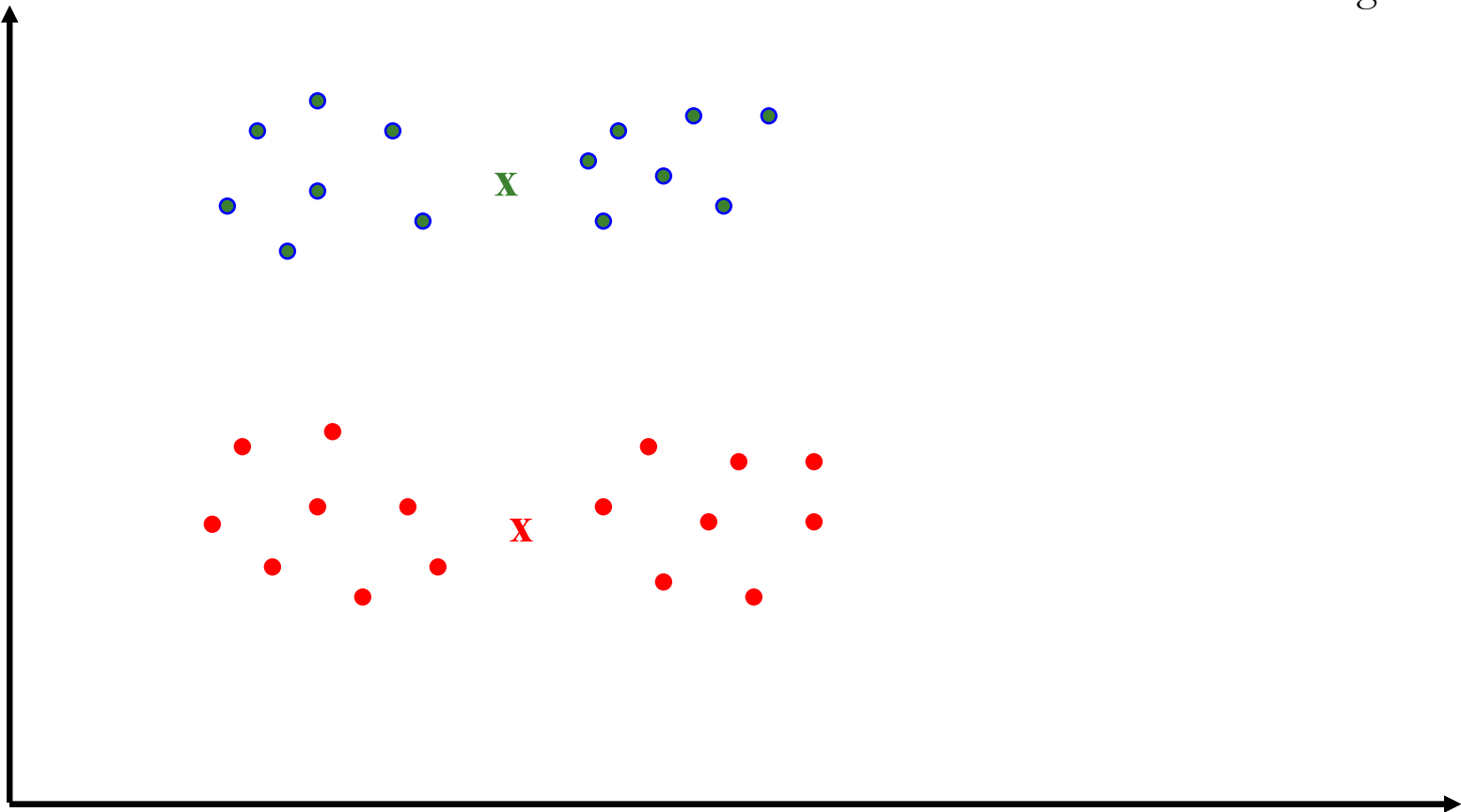


Clustering Algorithms Used for Validation: K-Means Illustration



Clustering Algorithms Used for Validation: K-Means Illustration

Re-estimate Means and Converge



Clustering Algorithms Used for Validation: K-Means

■ Algorithm Outline

- Initialize K cluster centers μ_i randomly. $1 \leq i \leq K$
 - Repeat until *convergence*:
 - **Cluster Assignment Step:** Assign each data point x to the cluster X_j such that “ L_2 ” / “ L_1 ” distance of x from μ_i (center of X_j) is minimum
 - **Center Re-estimation Step:** Re-estimate each cluster center μ_i as the mean of the points in that cluster
-

Clustering Algorithms for dynamic 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
 - Difficult to rely on visual perception due to noise and high dimension
 - How appropriate is the clustering method for the data at hand?
 - Would it be possible to have a better set of clusters?
 - Trade-off between computational cost and cluster quality
 - How much tolerance is accepted?
-

Clustering of PET data: Average Computational Costs (Minutes) (2-D, slice 16)

#Clusters	3	4	5	6	7
HAL	(128,191)	(122,187)	(122,185)	(120,186)	(113,186)
HAL1	(2.31,8.28)	(1.53,7.95)	(0.56,8.13)	(1.04,7.68)	(0.53,7.3)
HAL2	(2.06,8.28)	(1.29,6.89)	(1.22,6.11)	(1.58,6.23)	(1.20,5.59)
HCL	(124,185)	(122,184)	(122,184)	(120,184)	(113,183)
HCL1	(1.70,3.83)	(1.29,3.49)	(0.83,3.23)	(0.84,3.85)	(0.71,3.59)
HCL2	(0.44,1.43)	(0.52,0.65)	(0.27,0.47)	(0.20,0.40)	(0.19,0.41)
K Means	(0.12,0.16)	(0.18,0.20)	(0.16,0.19)	(0.28,0.33)	(0.34,0.79)

Cluster Validation

- Procedure of evaluating results of a clustering algorithm in a quantitative and objective manner
 - *Index* of cluster validity
 - Used to measure the quality and appropriateness of a clustering structure.
 - Approach employed in this study
 - Intra cluster characteristics (Compactness, low magnitude)
 - Inter cluster characteristics (Well-separated, high magnitude)
 - Measures based on combination of both (Overall assessment)
-

Cluster Validation Metrics:

Intra-Cluster

■ Average Distance from Mean

- For each element x_i in cluster j , $1 \leq j \leq p$

$$m_i^j = d(x_i, \mu_j) \text{ where } \mu_j = 1/n_j (\sum_m x_m^j)$$

- For each cluster

$$m^j = 1/n_j (\sum_i m_i^j)$$

- For all (p) clusters

$$m = 1/p \sum_j m^j$$

- Sensitive to noisy points
-

Cluster Validation Metrics:

Intra-Cluster

- Maximum Distance from Mean

$$r^j = \max_{1 \leq i \leq n_j} d(x_i, \mu_j)$$

- Measures distance from edge to center (radius); influences HCL algorithm
- Highly sensitive to noise

- Maximum Diameter

$$\Delta(X_j) = \max_{x, y \in X_j} d(x, y)$$

- Distance between the farthest points within a cluster
 - Tests for width of cluster; Sensitive to noisy points
-

Cluster Validation Metrics:

Intra-Cluster

■ Average Spread

$$a_i^j = \frac{1}{n_j} \sum_m d(x_i^j, x_m^j)$$

- Average distance of elements within a cluster to all other elements of the same cluster
- Measure of homogeneity of cluster

■ Total Energy

- Sum of squares of the distances to the mean for all cluster points (Recall: K-means objective function)
 - Expected to be least for K-means
 - Useful measure to compare hierarchical methods with K-means
-

Cluster Validation Metrics:

Inter-Cluster

■ Separation

- For each element : Average distance of element i in cluster j to elements of cluster k . ($k \neq j$)
- For each cluster: Average of individual separation values
- Measure of average distance of separation

■ Minimum Separation

- min (separation values for a cluster)
 - Sensitive to noisy points
 - Useful to detect questionable cluster assignments
-

Cluster Validation Metrics:

Inter-Cluster

■ Average Split

$$s_i^{jk} = \min_{x_i^j \in X_j, x_m^k \in X_k} d(x_i^j, x_m^k)$$

- Split: Closest point in cluster k to point i in cluster j .
 - Average of split values for all points in a cluster; average for all clusters
 - How close are two neighboring clusters?
-

Cluster Validation Metrics:

Silhouette

- Measures the *standardized* difference between $b(i)$ and $a(i)$
- $a(i)$: average dissimilarity of object i to all other objects in its own cluster (average spread)
- $b(i)$: average dissimilarity of object i to all other objects in its nearest cluster (separation)

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

$$C_i^j = \begin{cases} \text{close to 1, well classified} \\ \text{close to 0, not certain} \\ \text{close to -1, mis-classified} \end{cases}$$

Average Silhouette Width

= Average value over all clusters

$$\begin{cases} > 0.5 & \text{Good Classification} \\ < 0.2 & \text{Lack of cluster structure} \end{cases}$$

Useful to determine Optimal Number of clusters (Maximum Width)

Cluster Validation Metrics:

Modified Dunn's Index*

- General form of Dunn's Index for a partition matrix U

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

- Modified Dunn's Index (Average Linkage)

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

- Modified Dunn's Index (Complete Linkage)

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

*Bezdek, J. C. and Pal, N.R., *Some new indexes of cluster validity*, IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics 28 (1998), no. 3, 301-315.

Cluster Validation Metrics:

Modified Dunn's Index

- Modified Dunn's Index (Combined Average)

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

- All 3 measures
 - Useful for determining optimal number of clusters
 - Maximum value: Indicator of optimal number
 - Complete Linkage: Affected by noisy points
-

Results

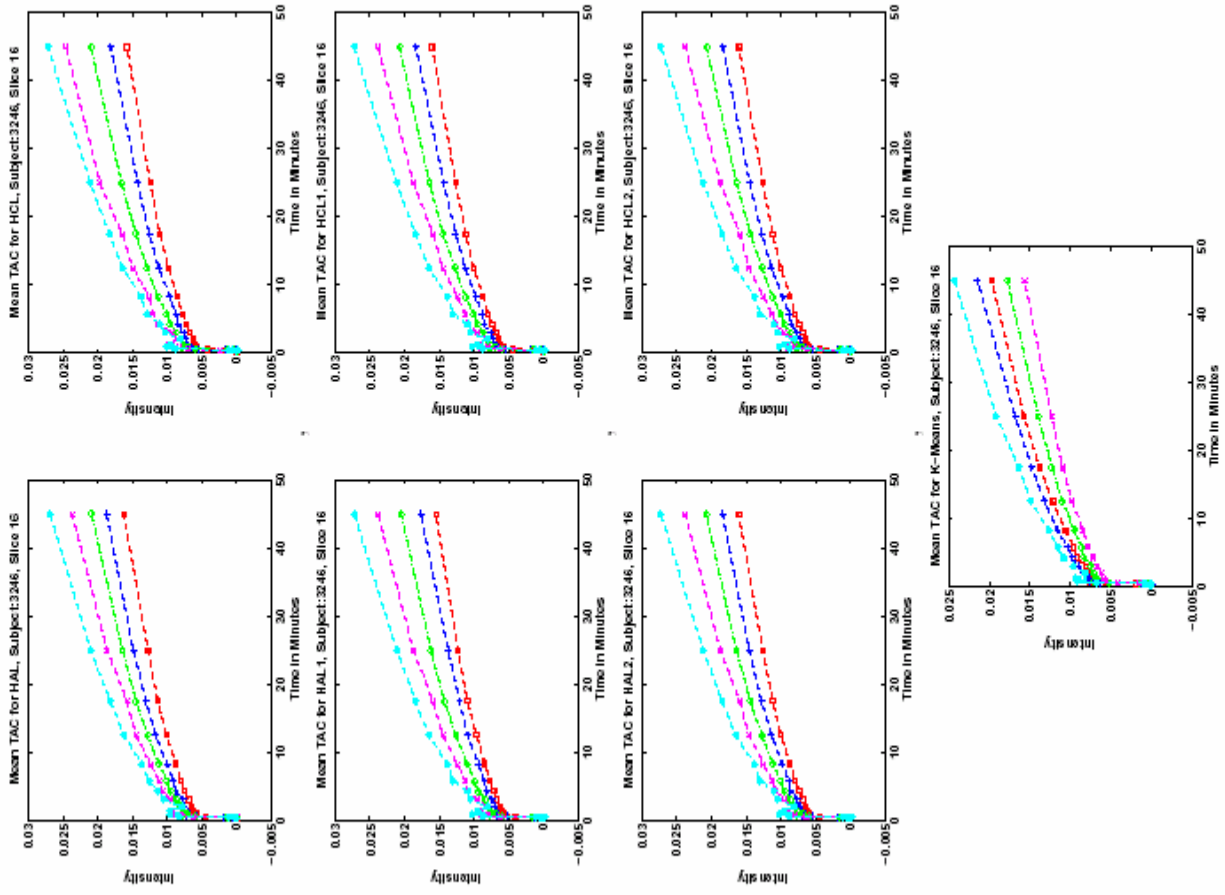


Figure 7. Mean TACs, All 7 Algorithms for subject 3246, slice=16, Number of Clusters=5

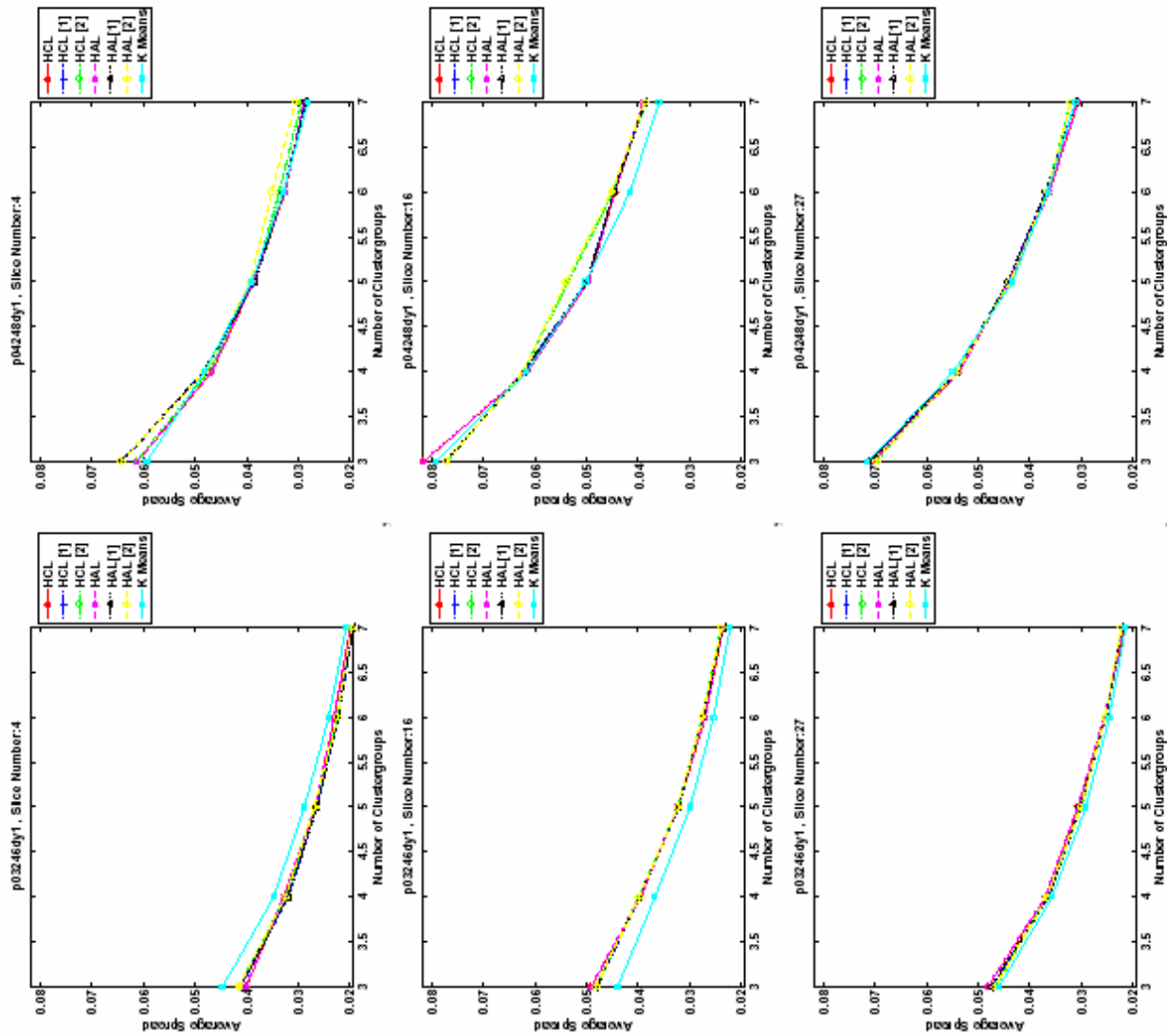


Figure 8. Average Distance to the mean Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

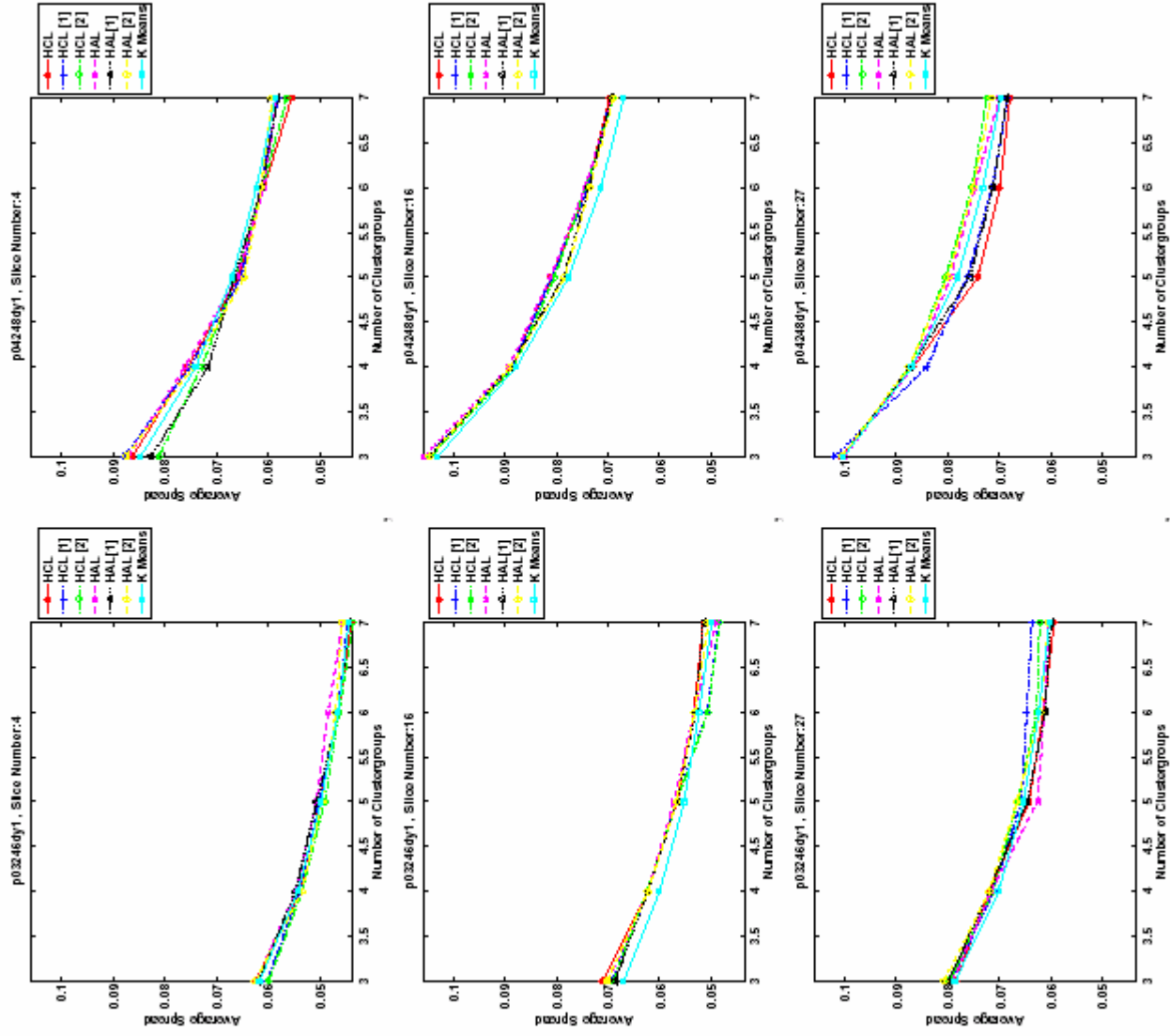


Figure 18. Average Distance to Mean Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

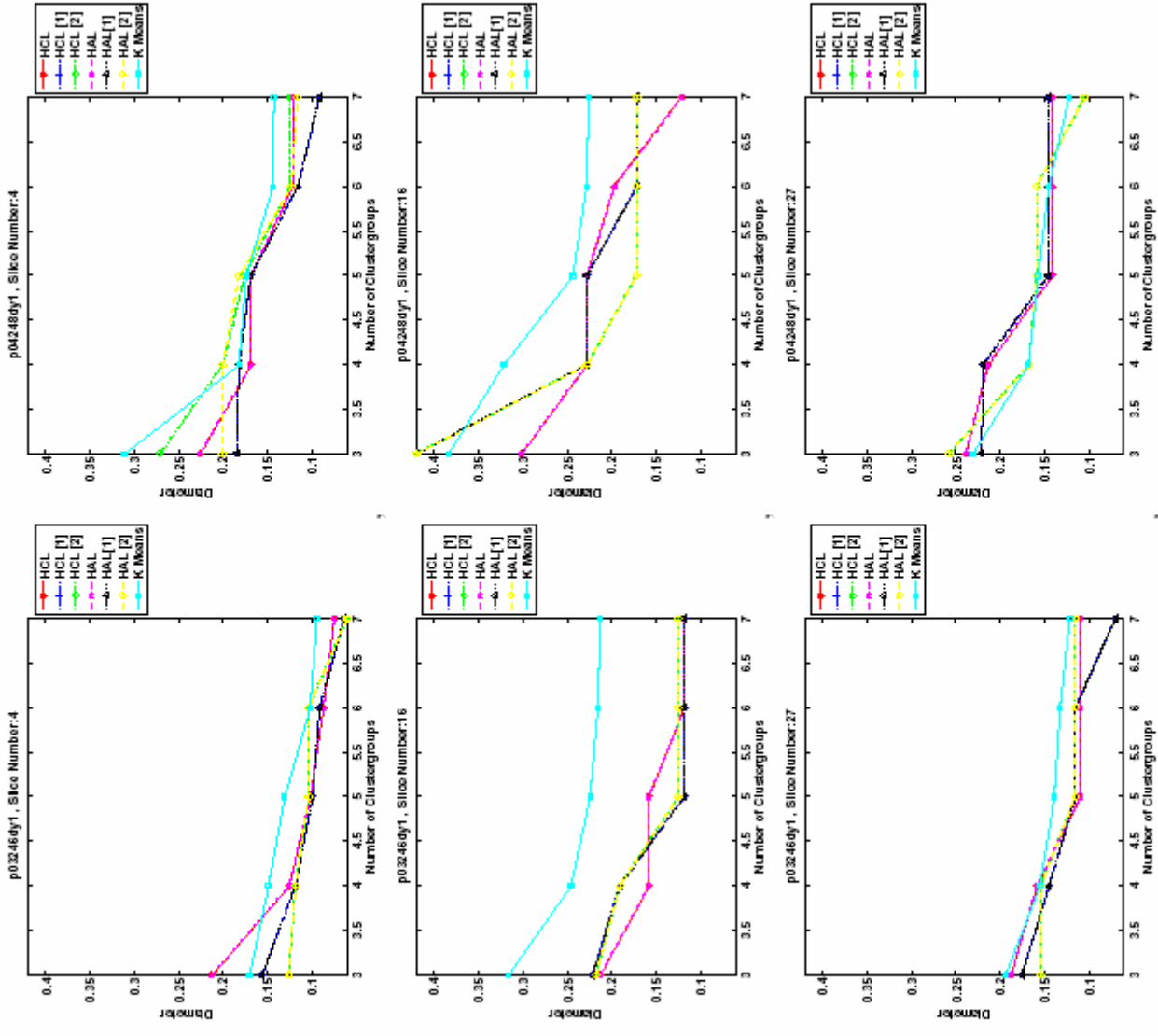


Figure 9. Maximum Distance to the Mean Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

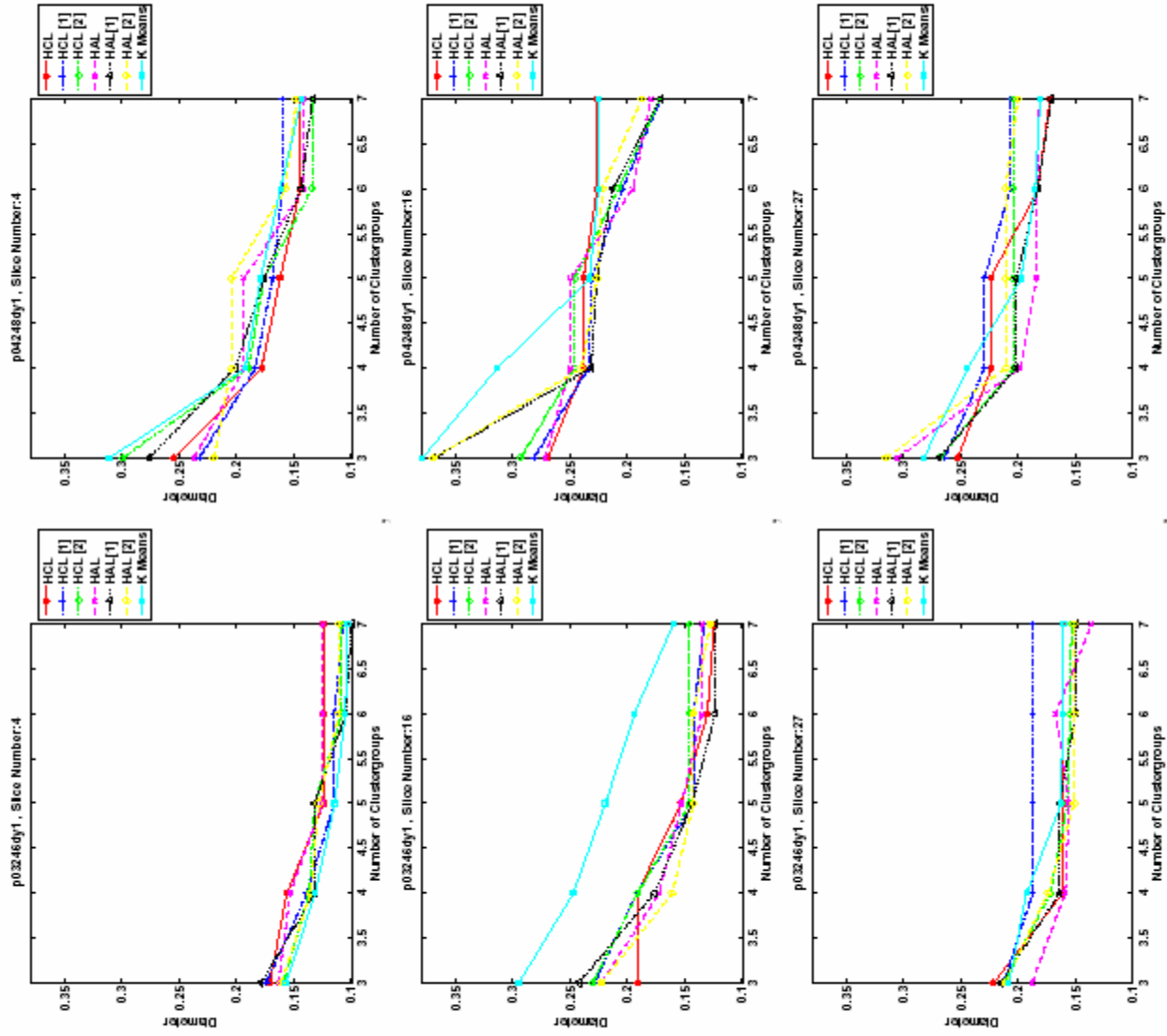


Figure 19. Maximum Distance to the Mean Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

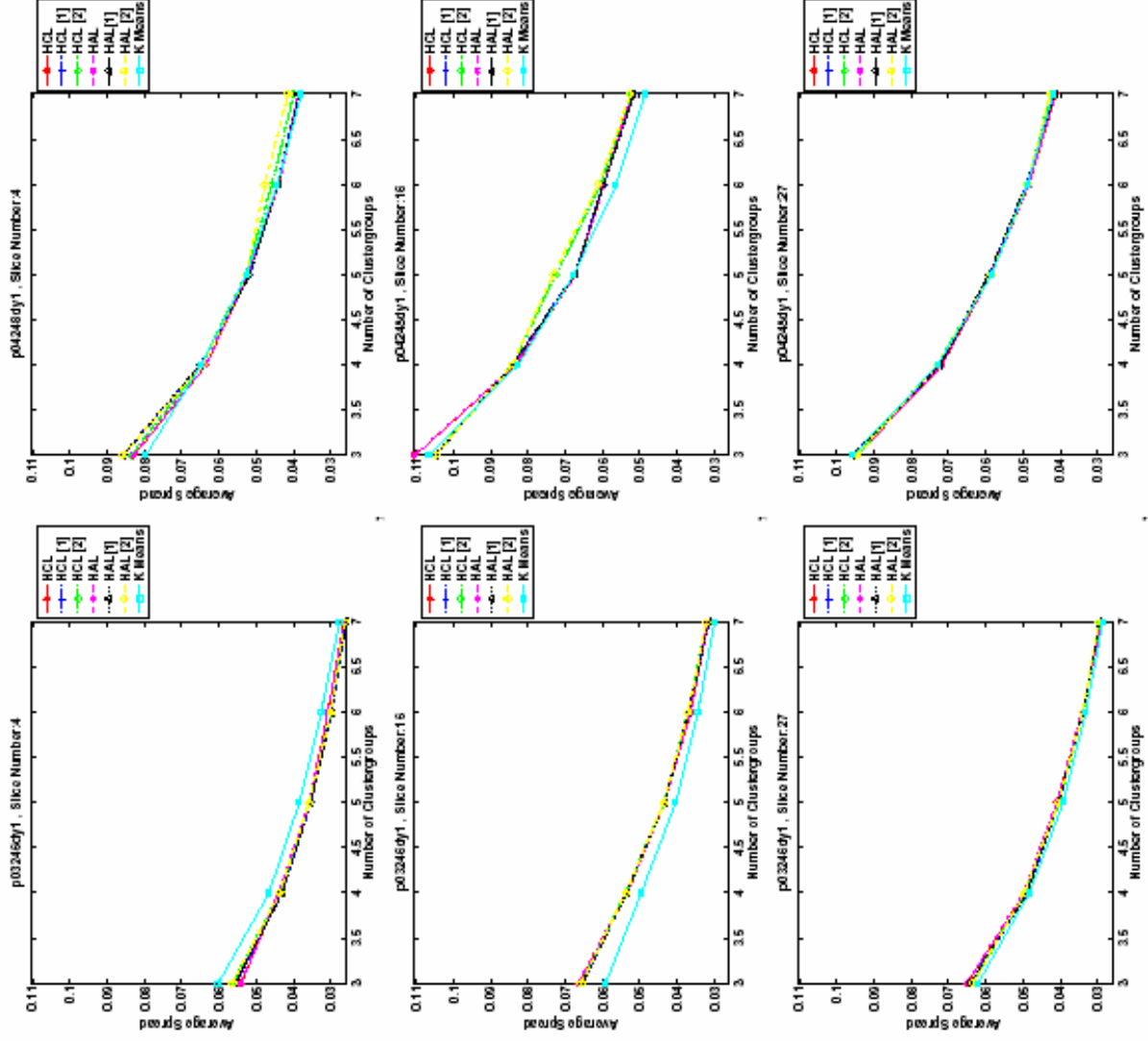


Figure 9. Average Spread Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

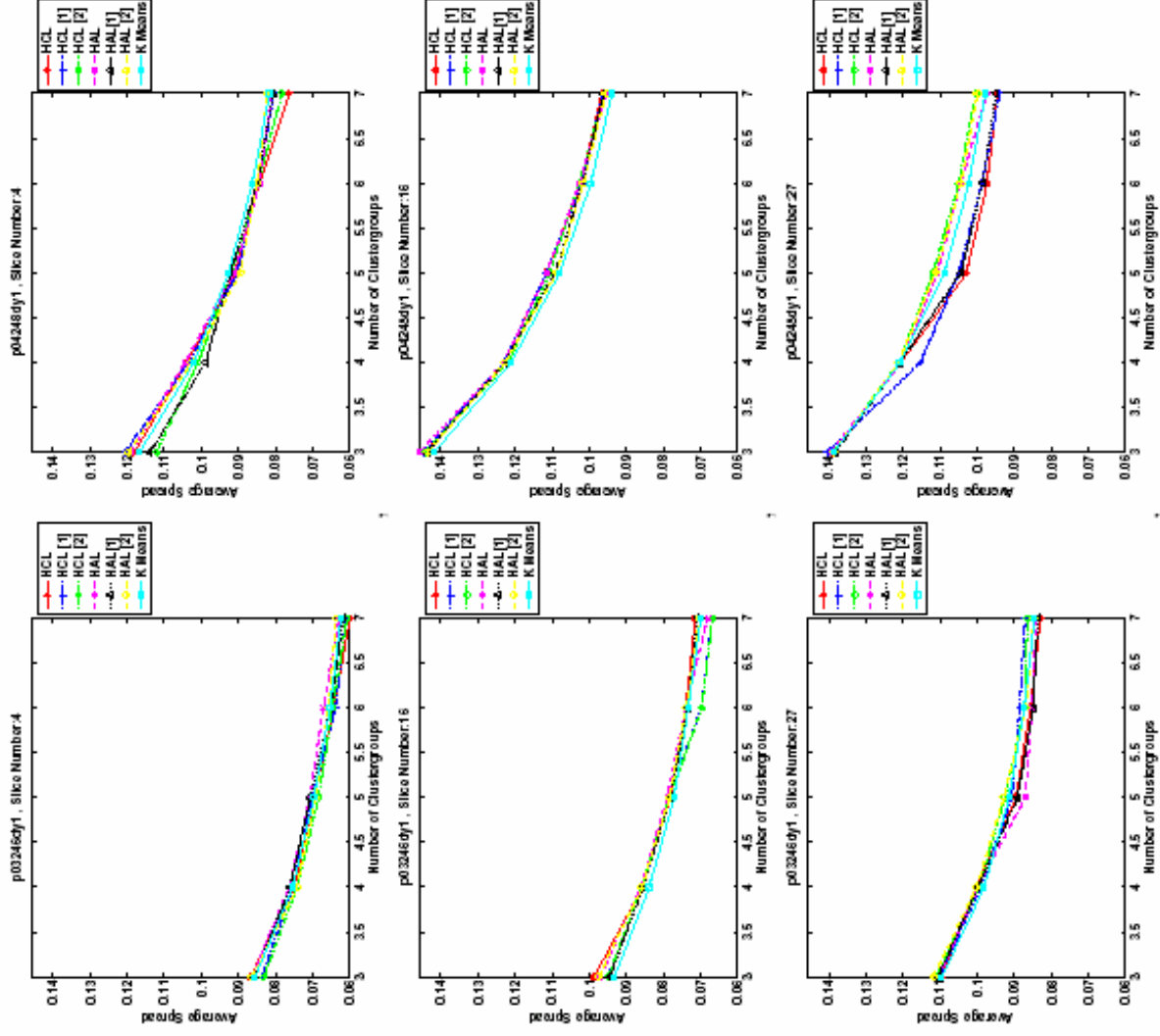


Figure 21. Average Spread Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

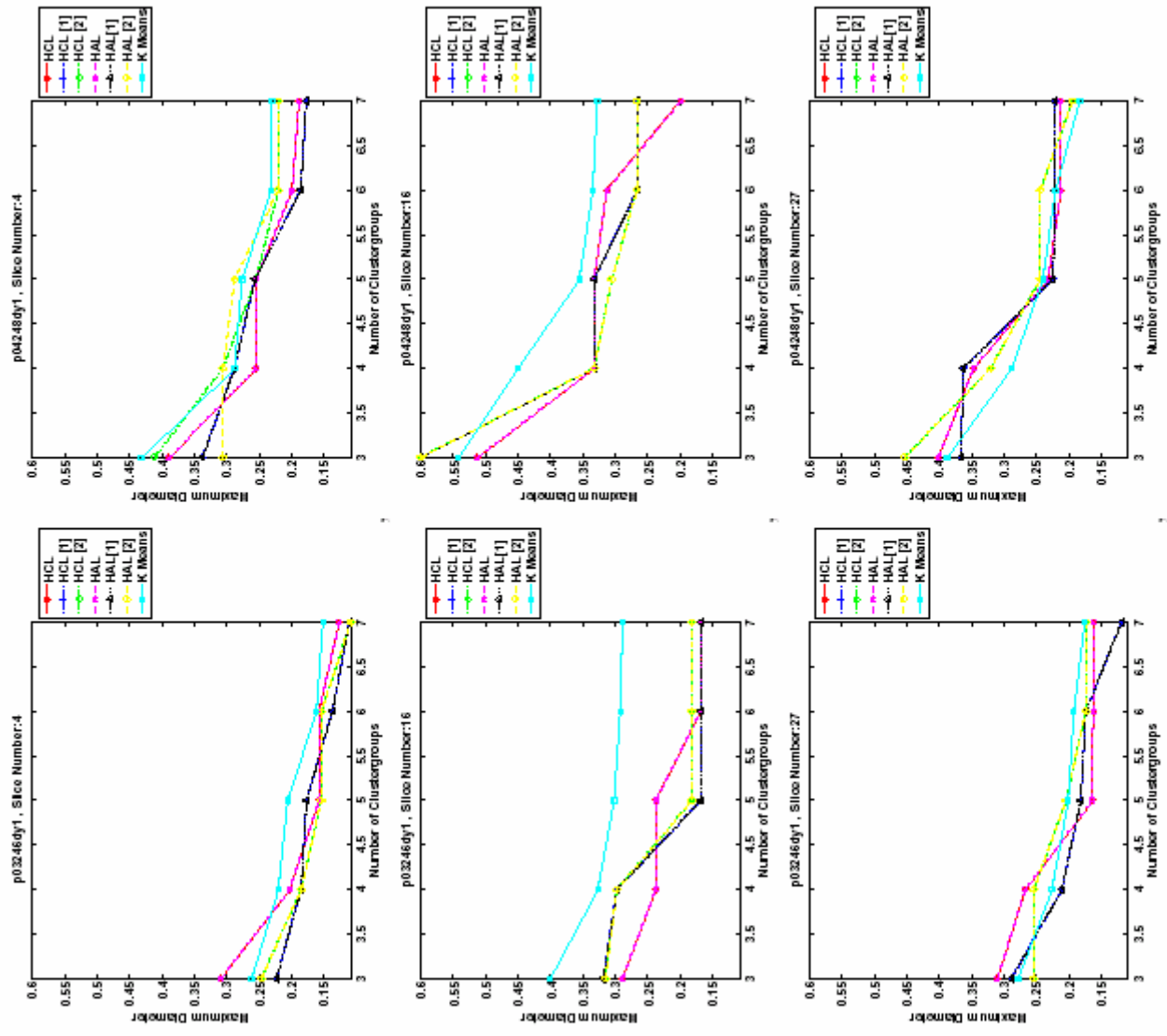


Figure 10. Maximum Diameter Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

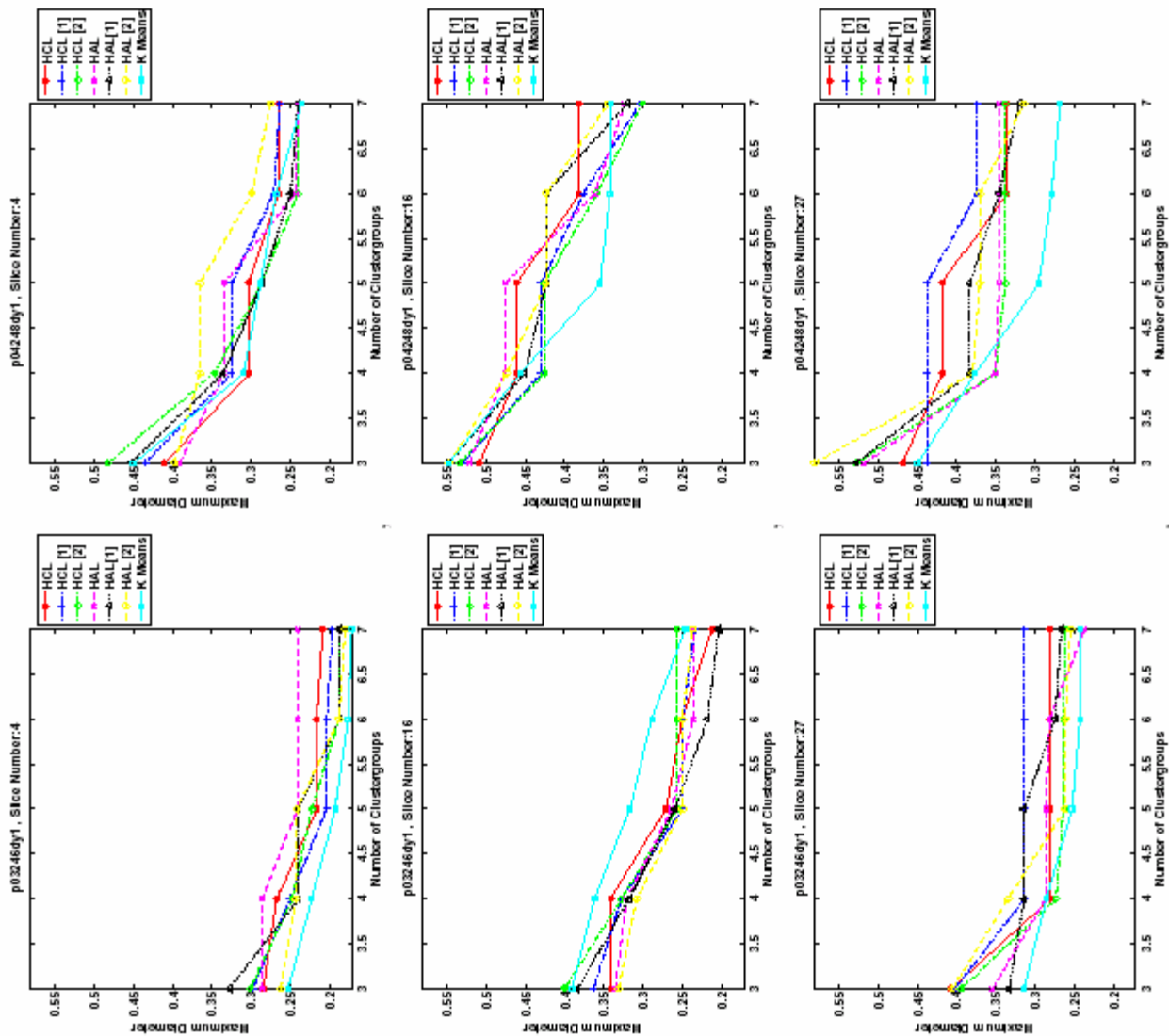


Figure 20. Maximum Diameter Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

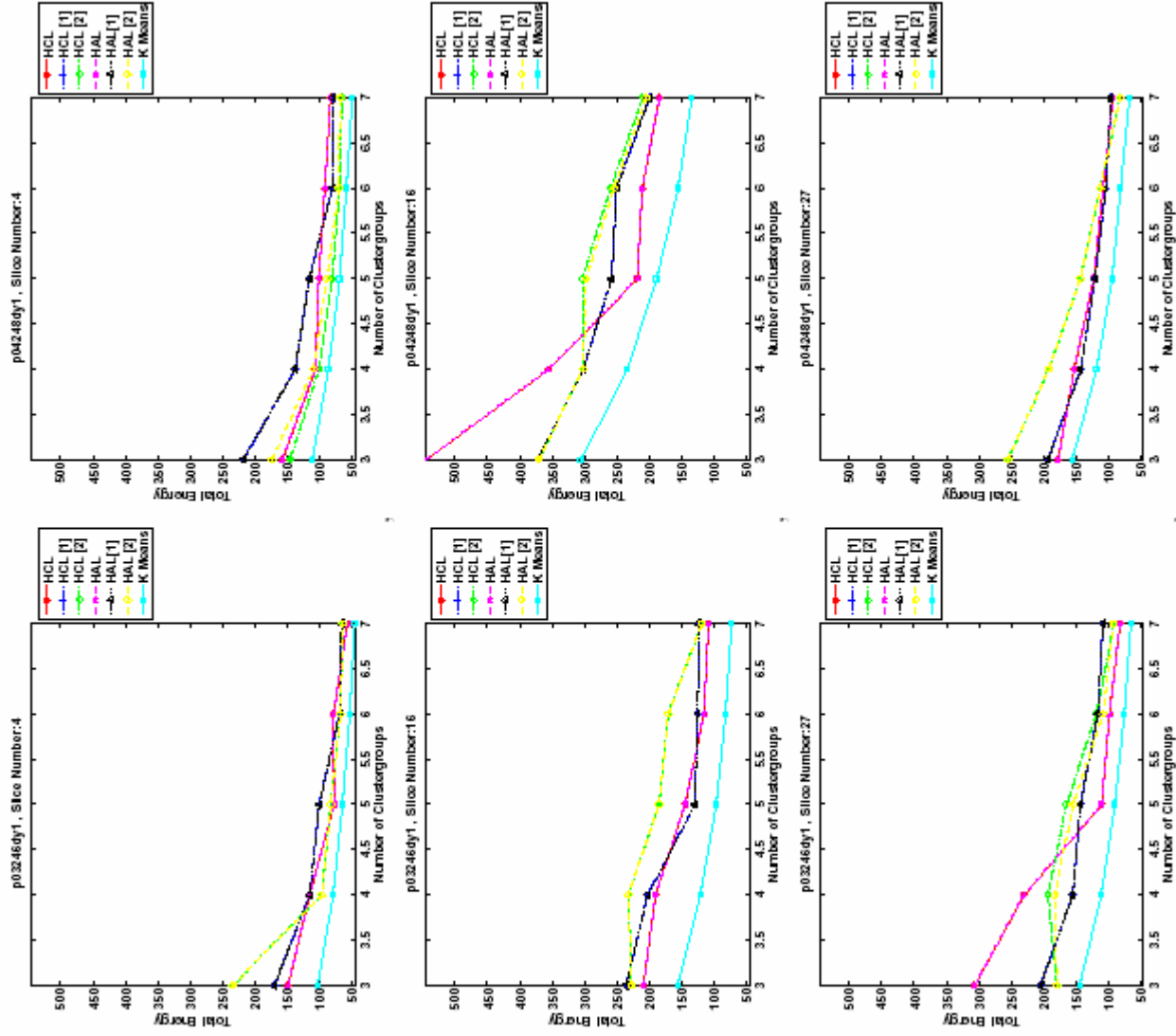


Figure 13. Total Energy Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

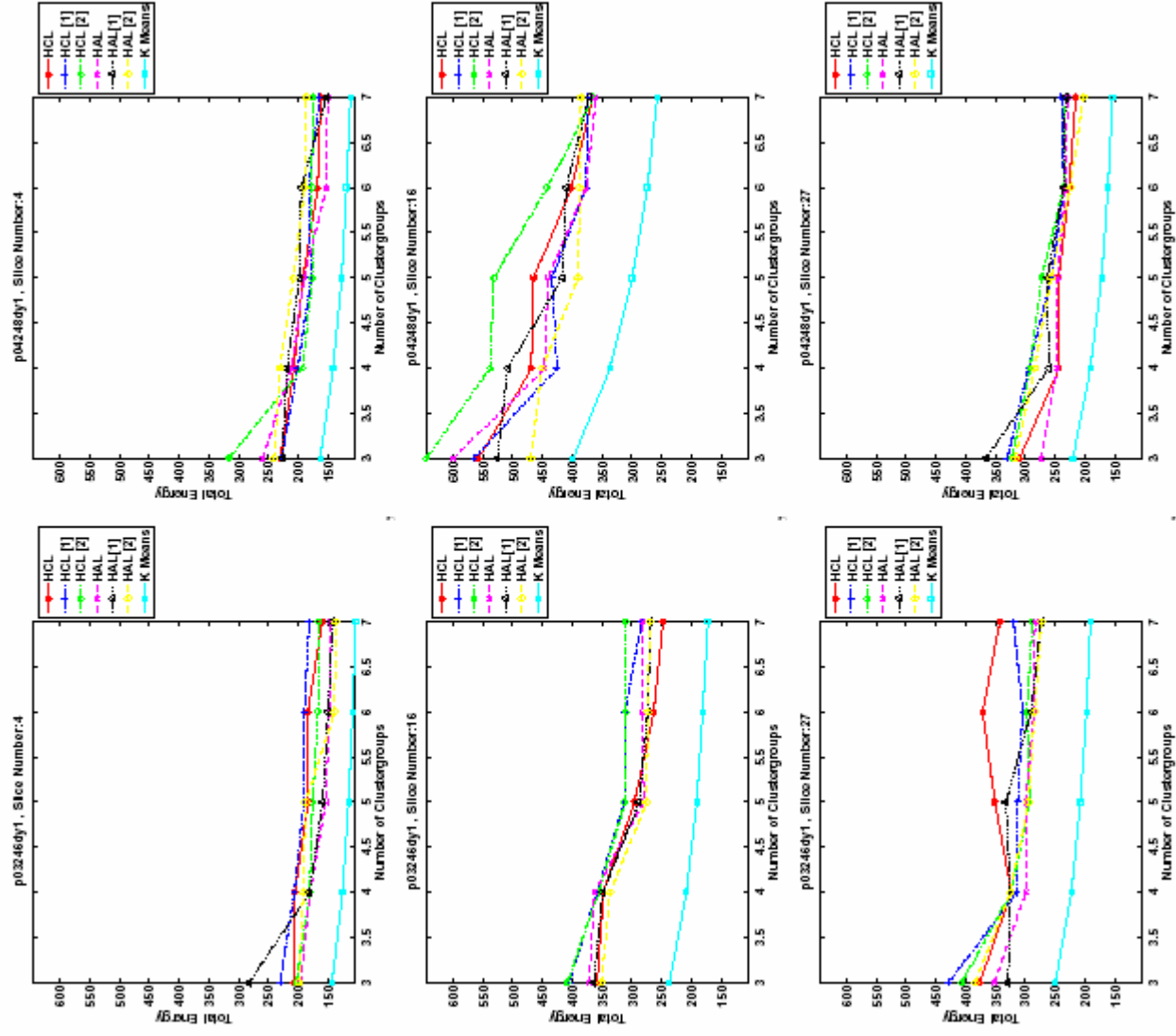


Figure 23. Total Energy Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

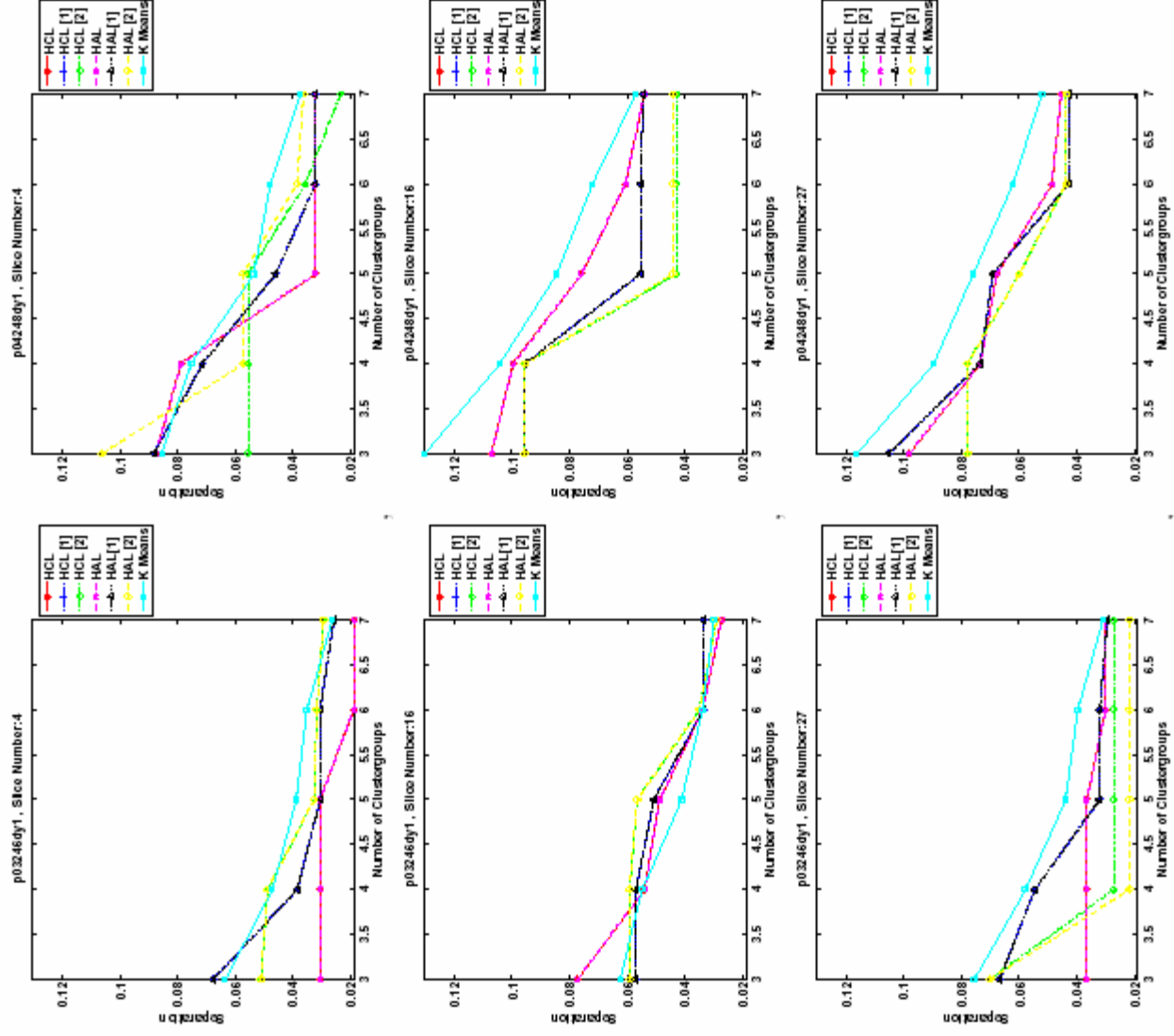


Figure 11. Minimum Separation Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

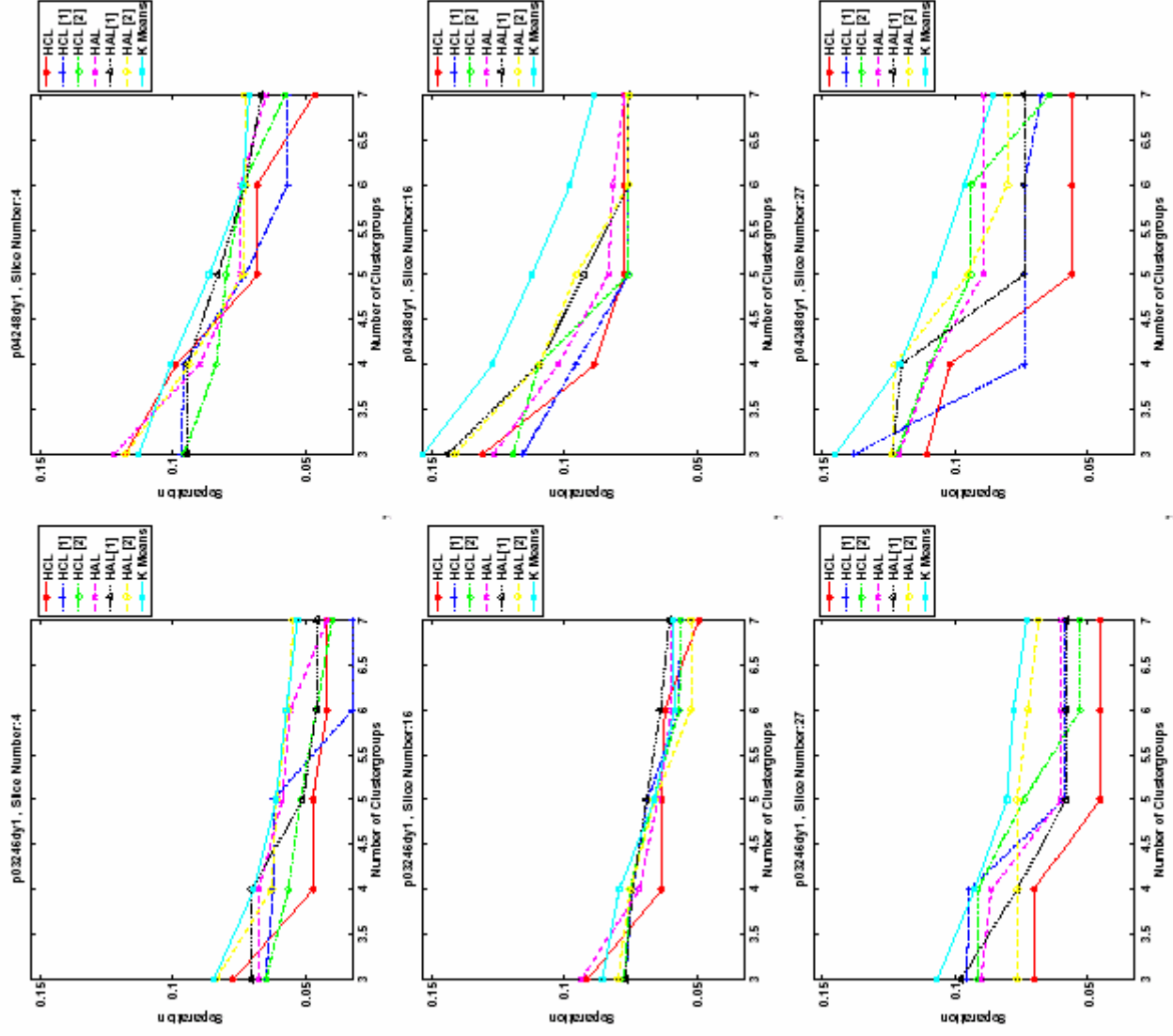


Figure 21. Minimum Separation Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

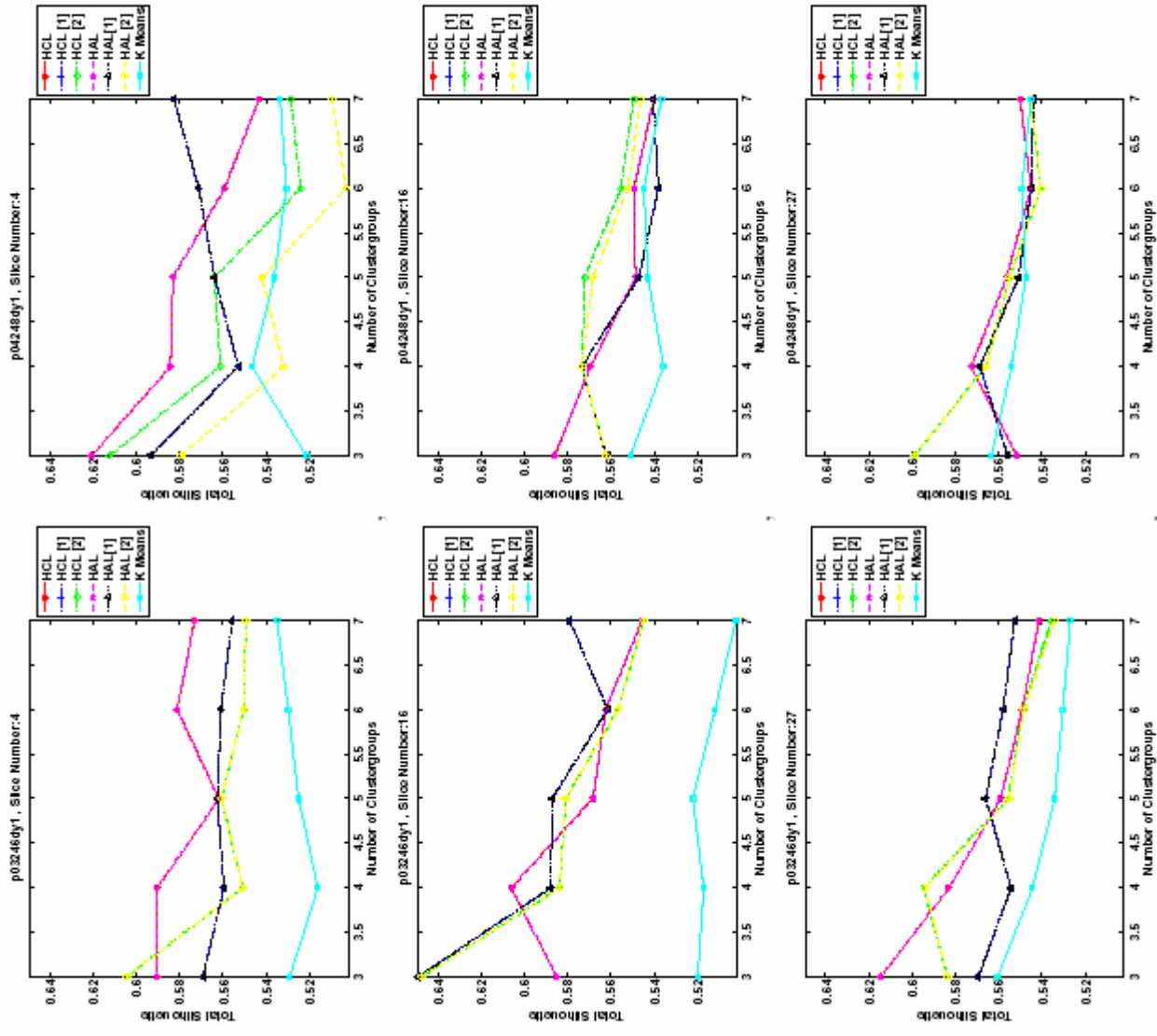


Figure 14. Silhouette Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

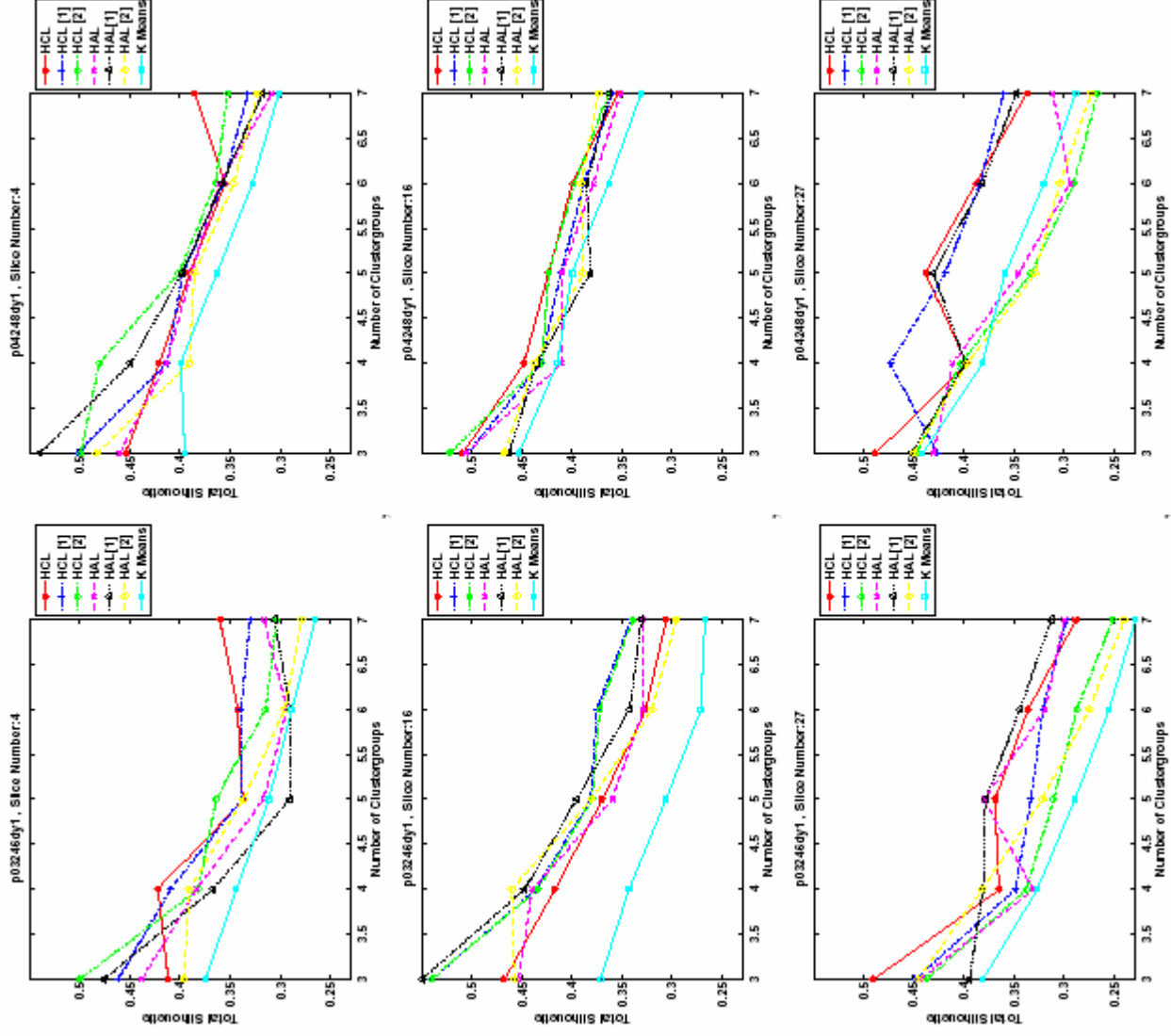


Figure 24. Silhouette Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

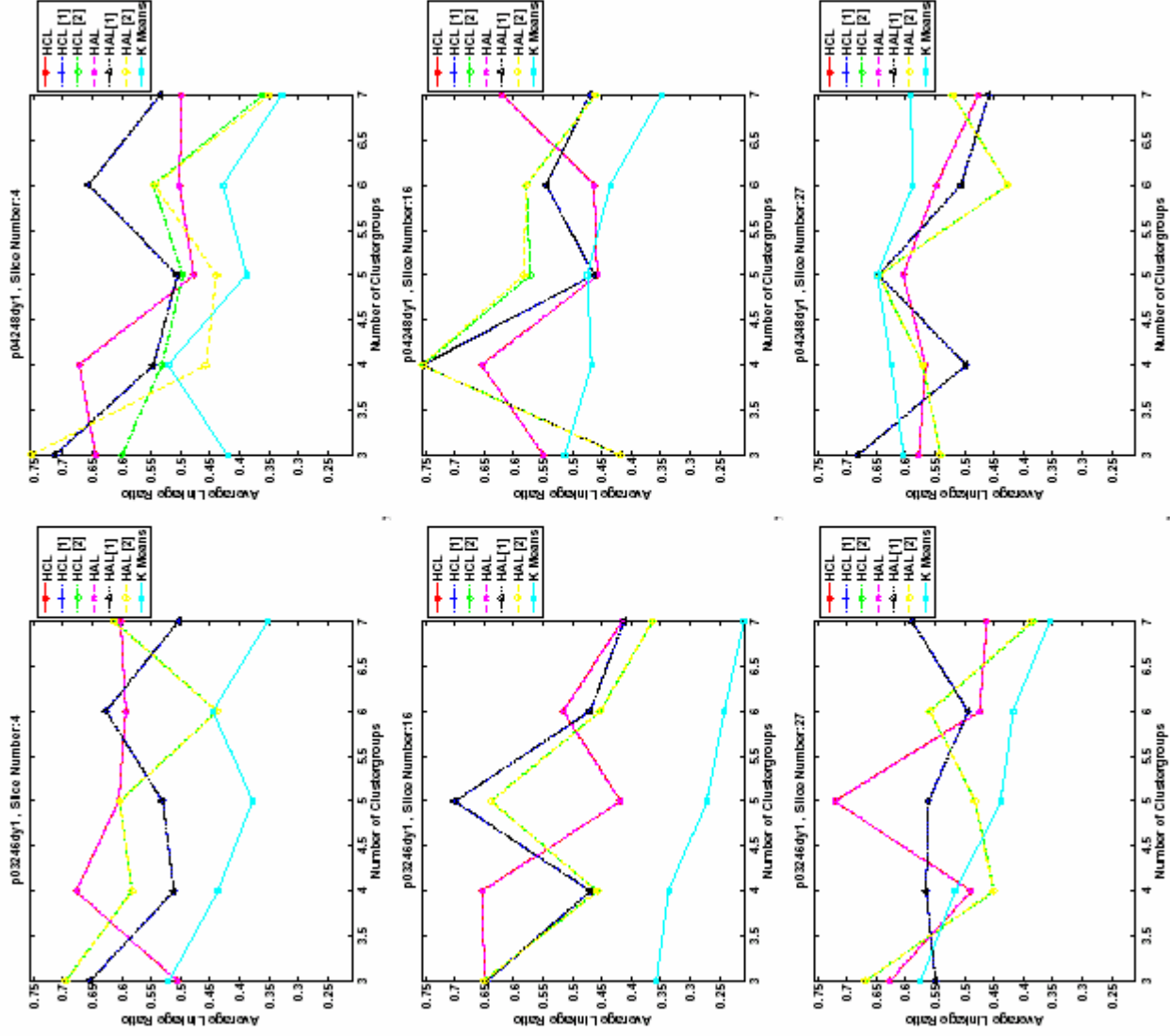


Figure 15. Average Linkage Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

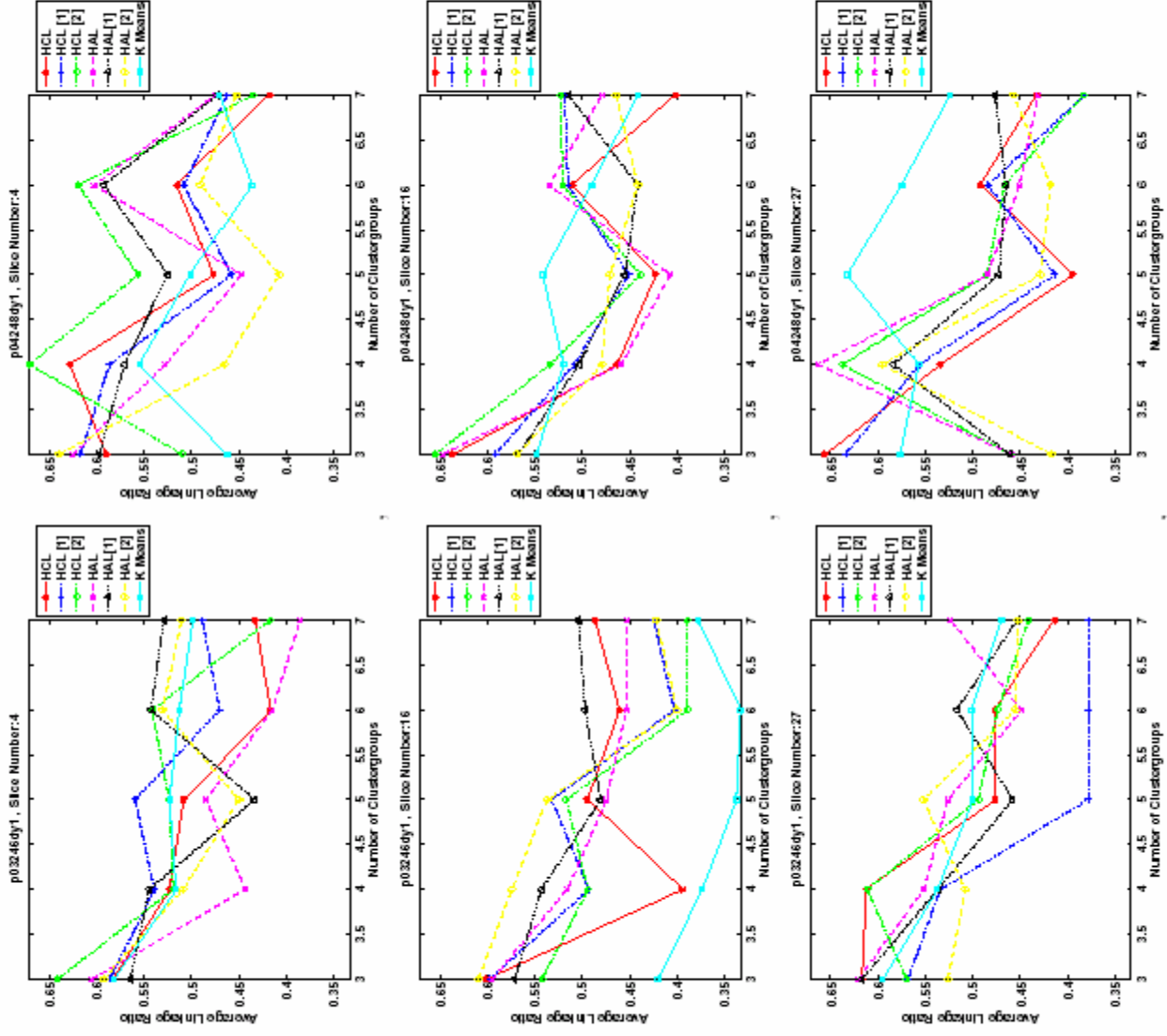


Figure 25. Average Linkage Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

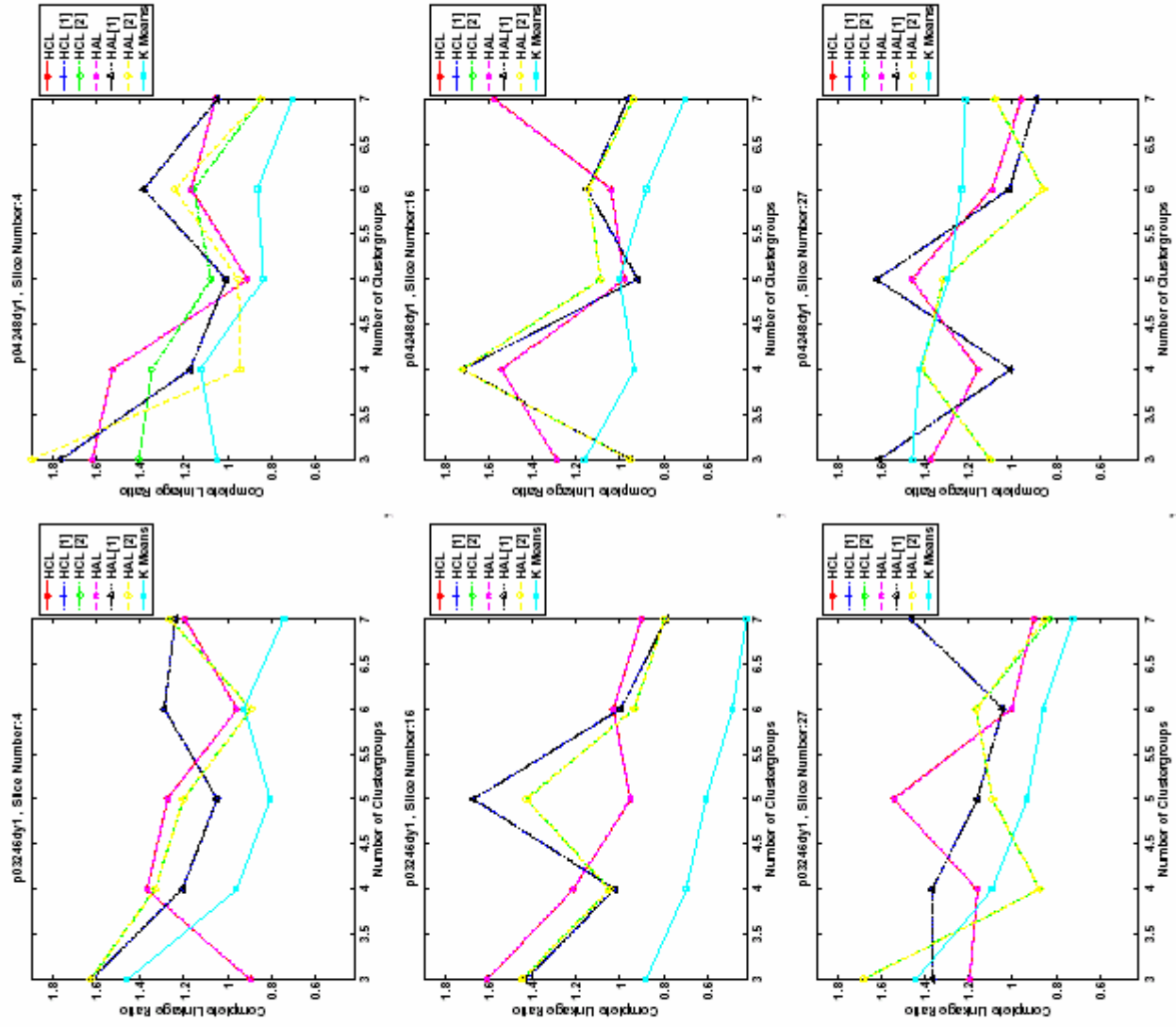


Figure 16. Complete Linkage Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

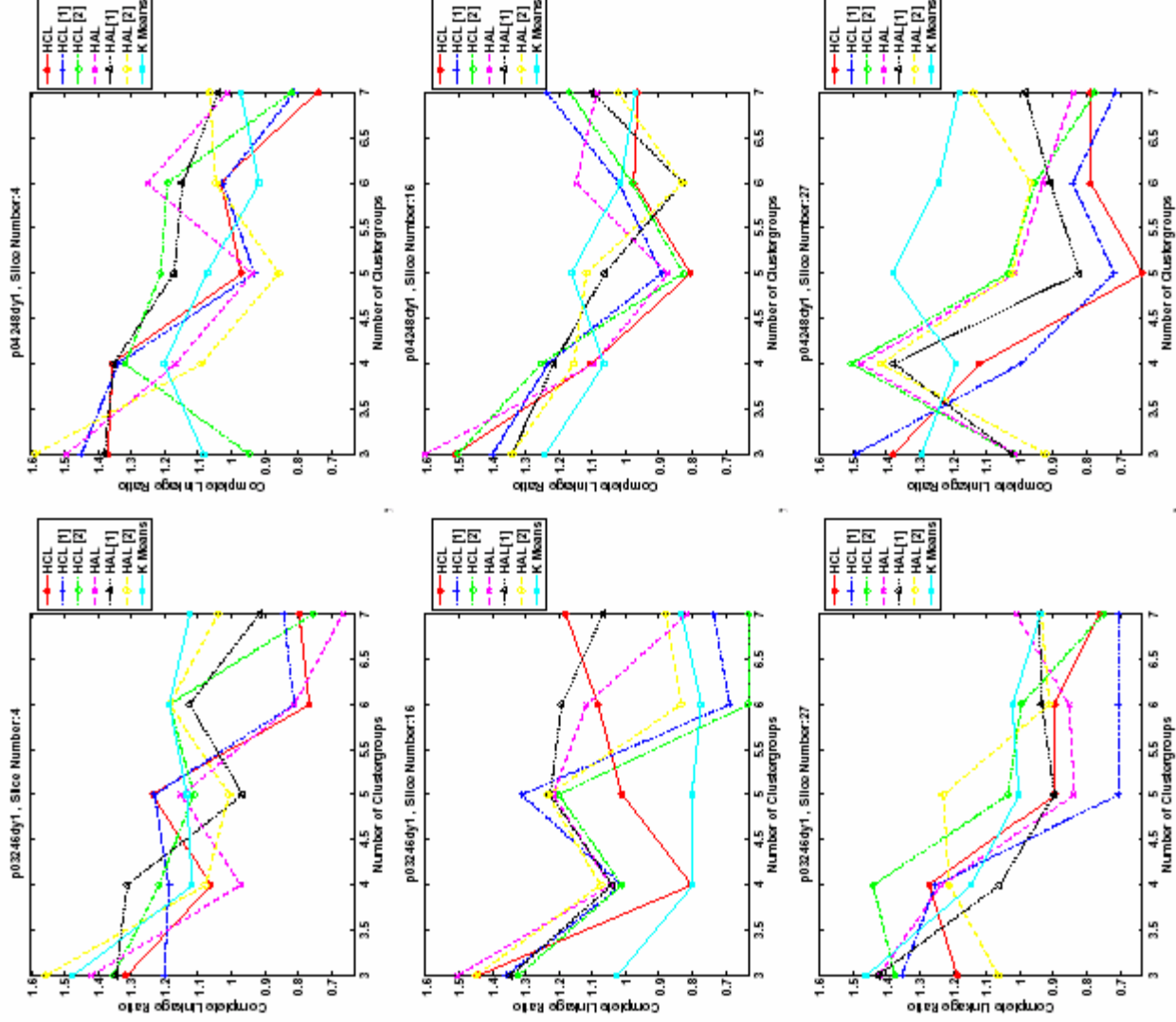


Figure 26. Complete Linkage Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

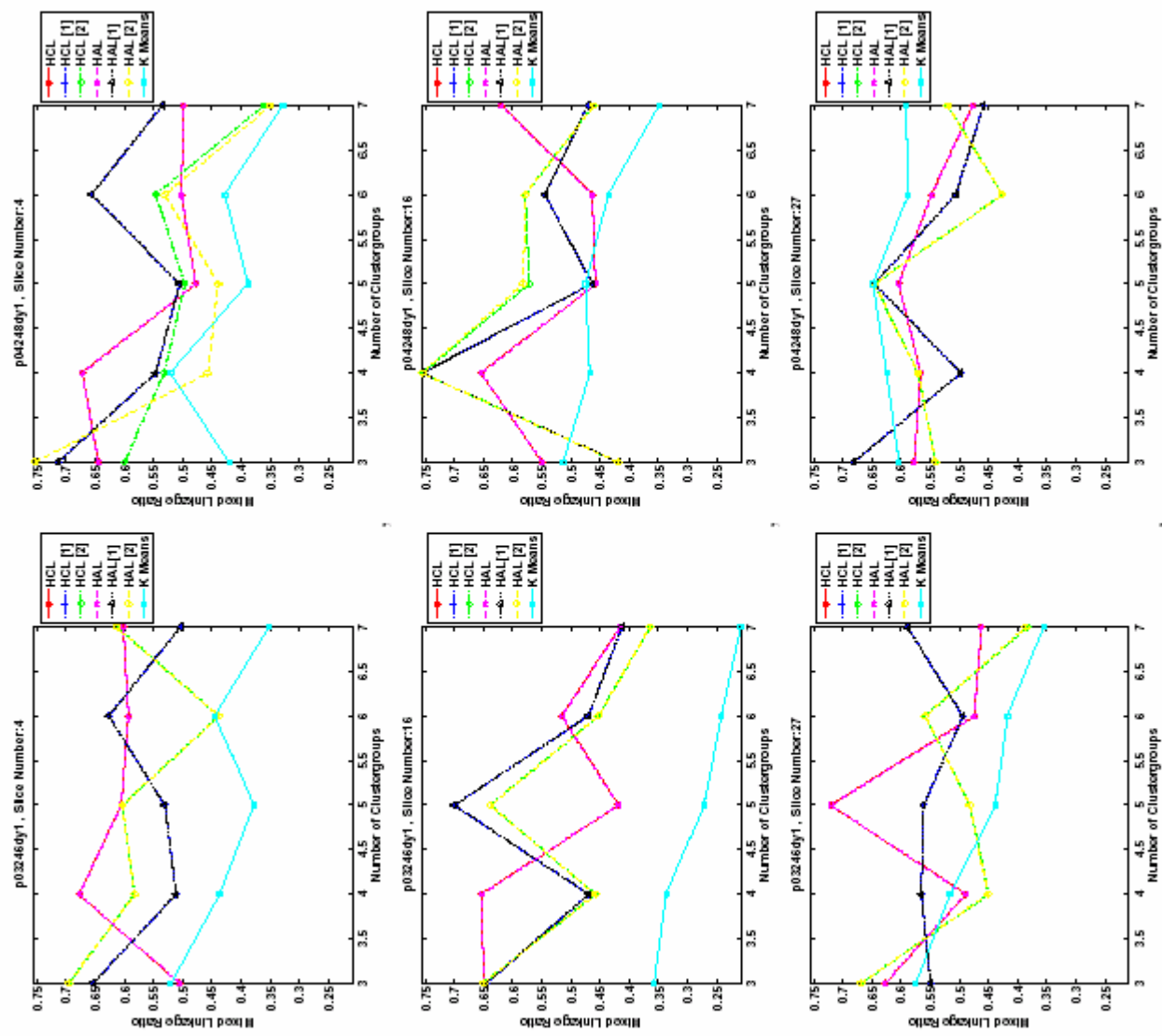


Figure 17. Combined Average Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

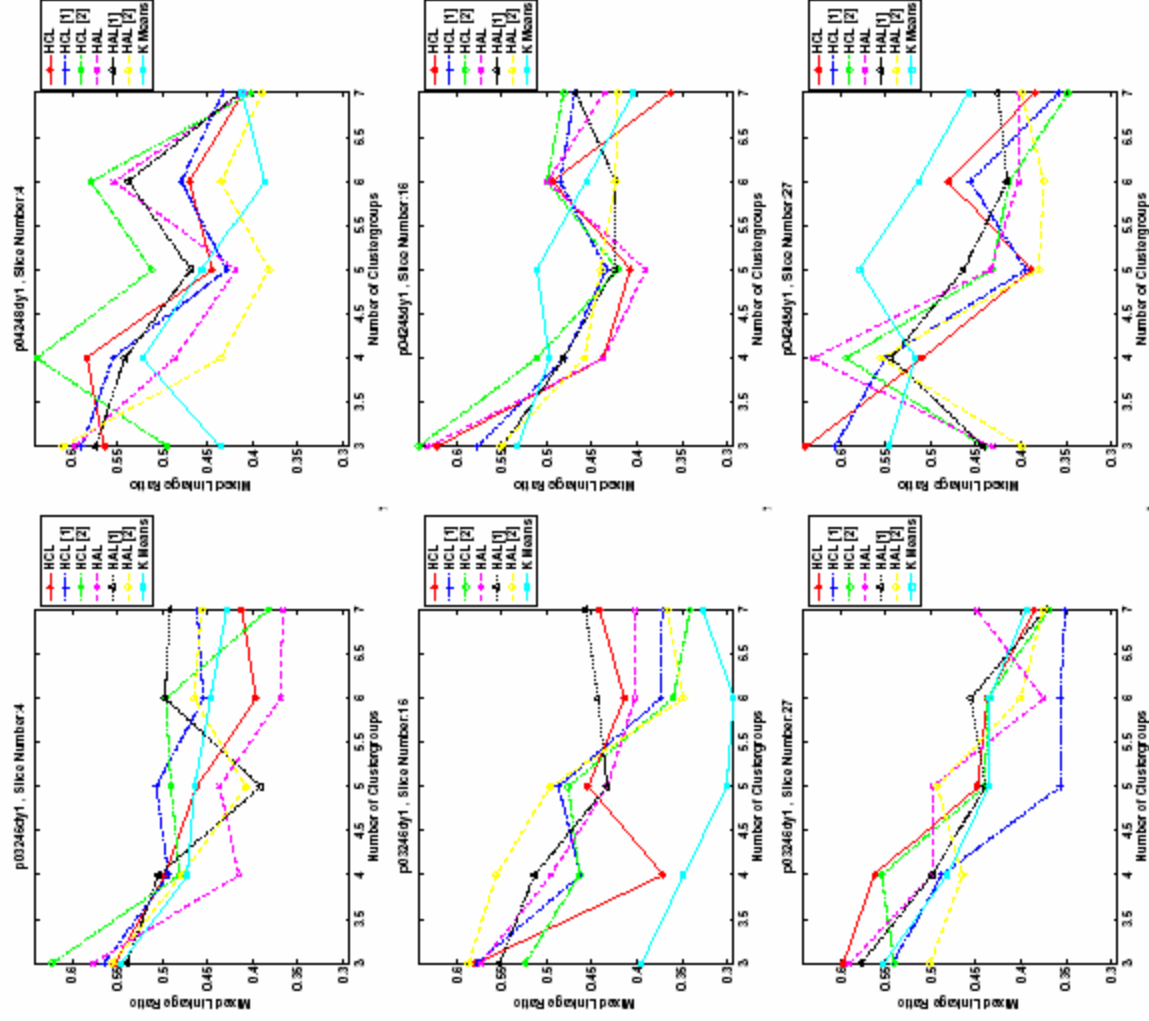


Figure 31. Combined Average Ratio Comparison Plots for slices 4, 16 and 27 of subjects #3246 and #4248

Conclusions

- Methods comparable to each other
 - Fast Hierarchical methods comparable to more expensive traditional methods
 - K-means
 - Greater average dissimilarity within clusters
 - Fast
 - Modestly well separated clusters
 - Optimal Number of clusters
 - Not very conclusive based on given measures
 - Most instances give value between 3 and 6.
 - Require domain expert knowledge to determine biological relevance
-

Future Work

- Fuzzy Clustering schemes, SOM, PCA
 - Using fast methods to exclusively study entire brain volume (in progress)
 - Advanced statistical techniques
 - More indices to obtain optimal number of clusters (Davies-Bouldin, Modified Hubert's Gamma statistic)
 - Publish all results and scripts
-

Things I Learned

- Practical application and use of clustering concepts
 - Very broad field, several applications, active research area
 - Hone Matlab scripting skills
 - Opportunity to deal with real data and understand associated challenges
 - Patience- Invaluable in such studies
-

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
 - 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
 - A.K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
-

Acknowledgements

- Dr. Rosemary Renaut
 - Dr. Hongbin Guo
 - Dr. Huan Liu
-