Cluster Analysis: Identifying Parkinson's Disease Subtypes

Jesse Mu

Wednesday, June 10

1 Preprocessing

1.1 Dataset Description

951 subjects, 145 metrics, collected 15-4-2012 from Pablo Martinez ín. Only 19 features used for clustering and/or interpretation. 50 subjects with missing values of the features to be used in clustering (brought down to 901). Imputation may be a good idea later on.

1.2 Selected Features

Combination of non-motor scale (NMS) symptoms and standard motor symptoms.

Name	Type	Format	Description
nms_d1	byte	%8.0g	cardiovascular
nms_d2	byte	%8.0g	sleep/fatigue
nms_d3	byte	%8.0g	mood/cognition
nms_d4	byte	%8.0g	percep/hallucinations
nms_d5	byte	%8.0g	attention/memory
nms_d6	byte	%8.0g	gastrointestinal
nms_d7	byte	%8.0g	urinary
nms_d8	byte	%8.0g	sexual function
nms_d9	byte	%8.0g	miscellaneous
tremor	float	%9.0g	tremor
bradykin	float	%9.0g	bradykinesia[1]
rigidity	float	%9.0g	rigidity
axial	float	%9.0g	axial[2]
pigd	float	%9.0g	postural instability and gait difficulty

Table 1: Selected Features and Details

[1] Impaired ability to adjust the body's position.

[2] Issues affecting the middle of the body.

Name	μ	σ	min-max
nms_d1	1.73	3.35	0-24
nms_d2	8.75	8.70	0-48
nms_d3	8.68	11.55	0-60
nms_d4	1.64	3.86	0-33
nms_d5	5.42	7.43	0-36
nms_d6	5.53	6.79	0-36
nms_d7	8.08	8.94	0-36
nms_d8	3.52	5.97	0-24
nms_d9	7.13	7.79	0-48
tremor	2.59	2.58	0-12
bradykin	2.40	1.41	0-6
rigidity	2.24	1.36	0-6
axial	3.25	2.68	0-12
pigd	3.31	2.71	0-12

Table 2: Descriptive Statistics

1.3 Dimensionality Reduction: PCA

May not be useful? If we're trying to identify *clinically* relevant features, merging them may not be a good idea. Regardless, Figure 1 shows results of preliminary PCA.

Figure 2 shows scree test elbow occurs around 2 or 2 or .4 Also, eigenvalues $1 - 5 > 1$.

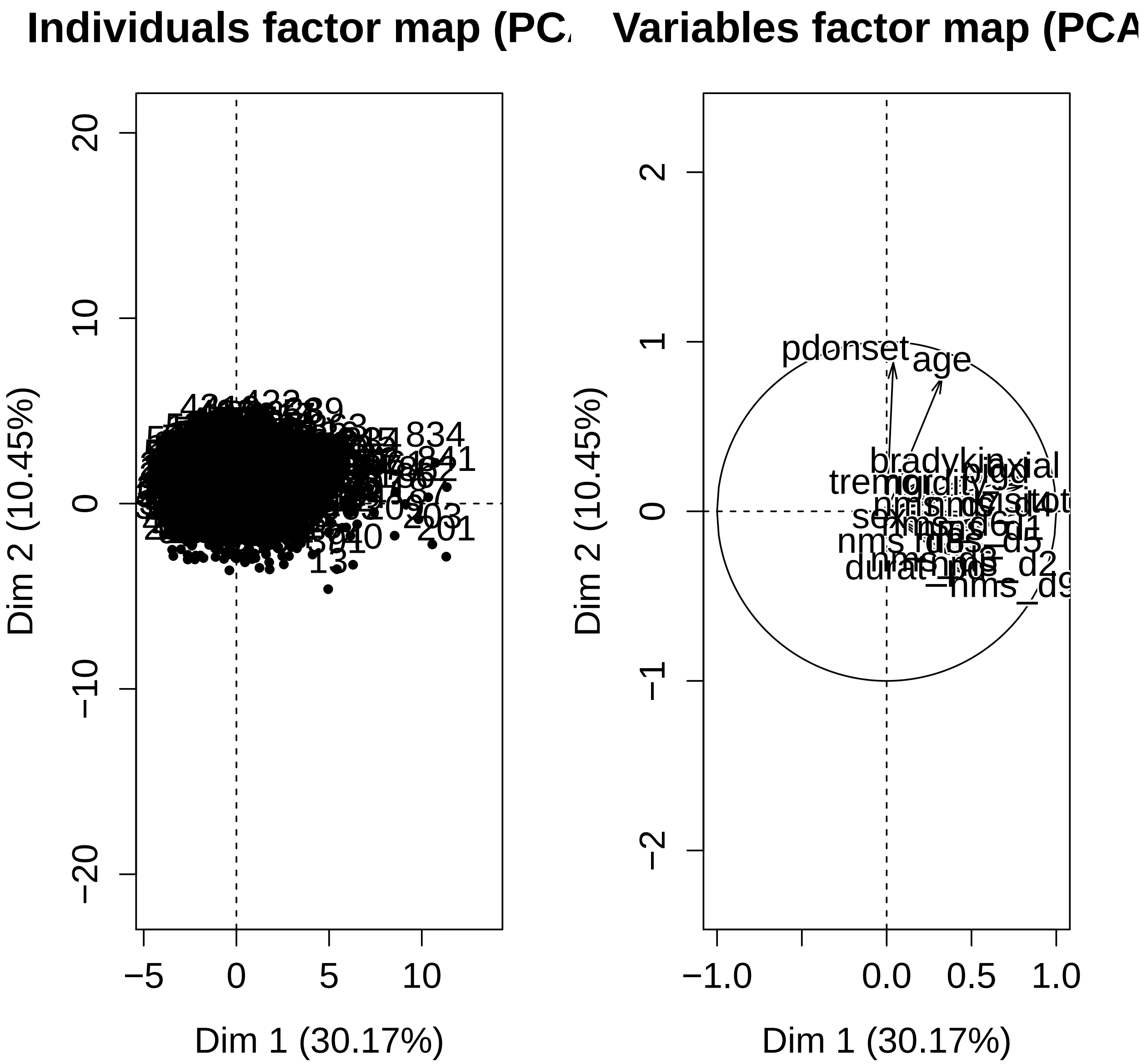

Figure 1: PCA Analysis

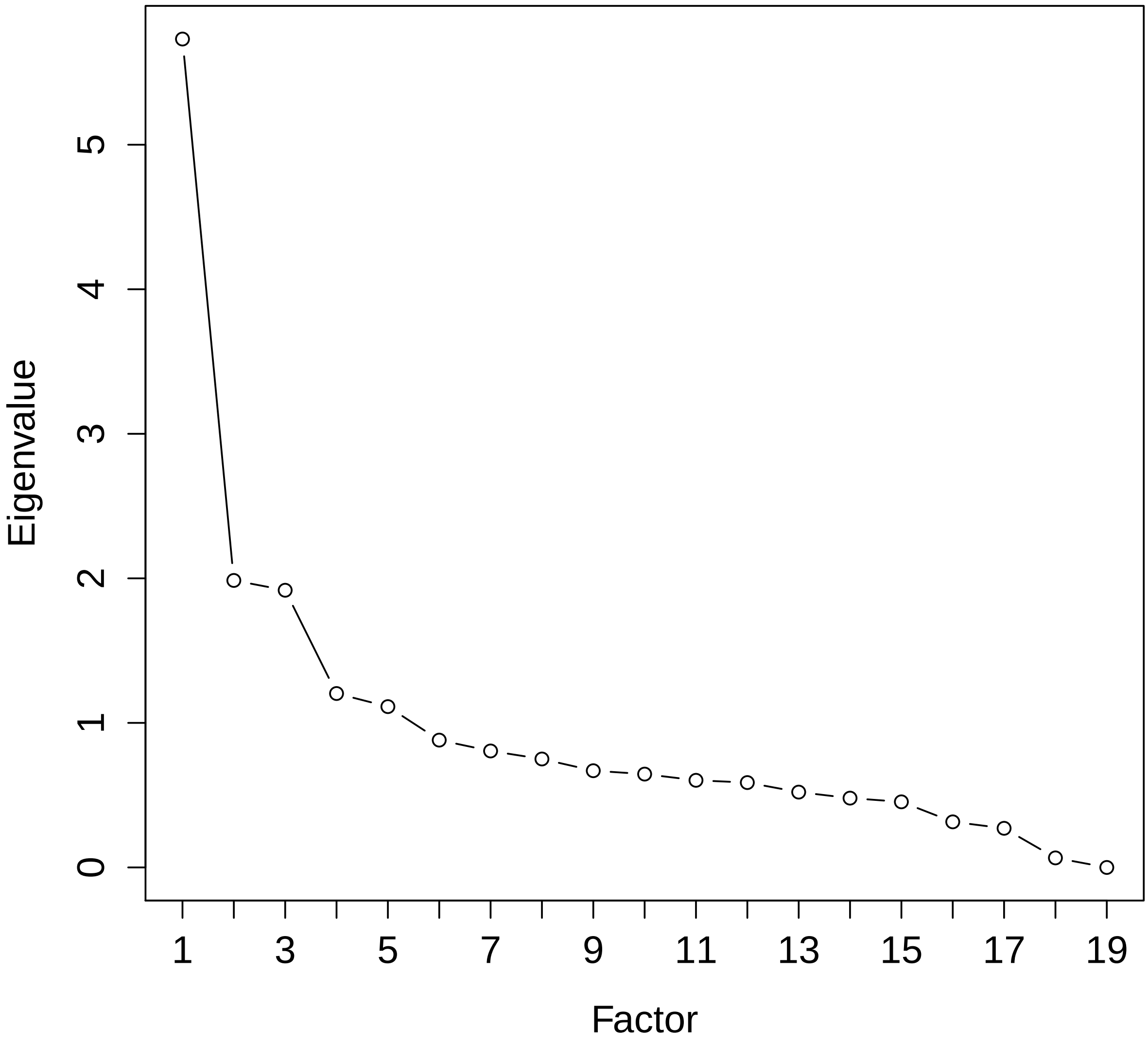

Figure 2: Scree test: eigenvalues by factor

2 k-means

2.1 Identifying optimal number of clusters

2.1.1 WSS Error Scree Test

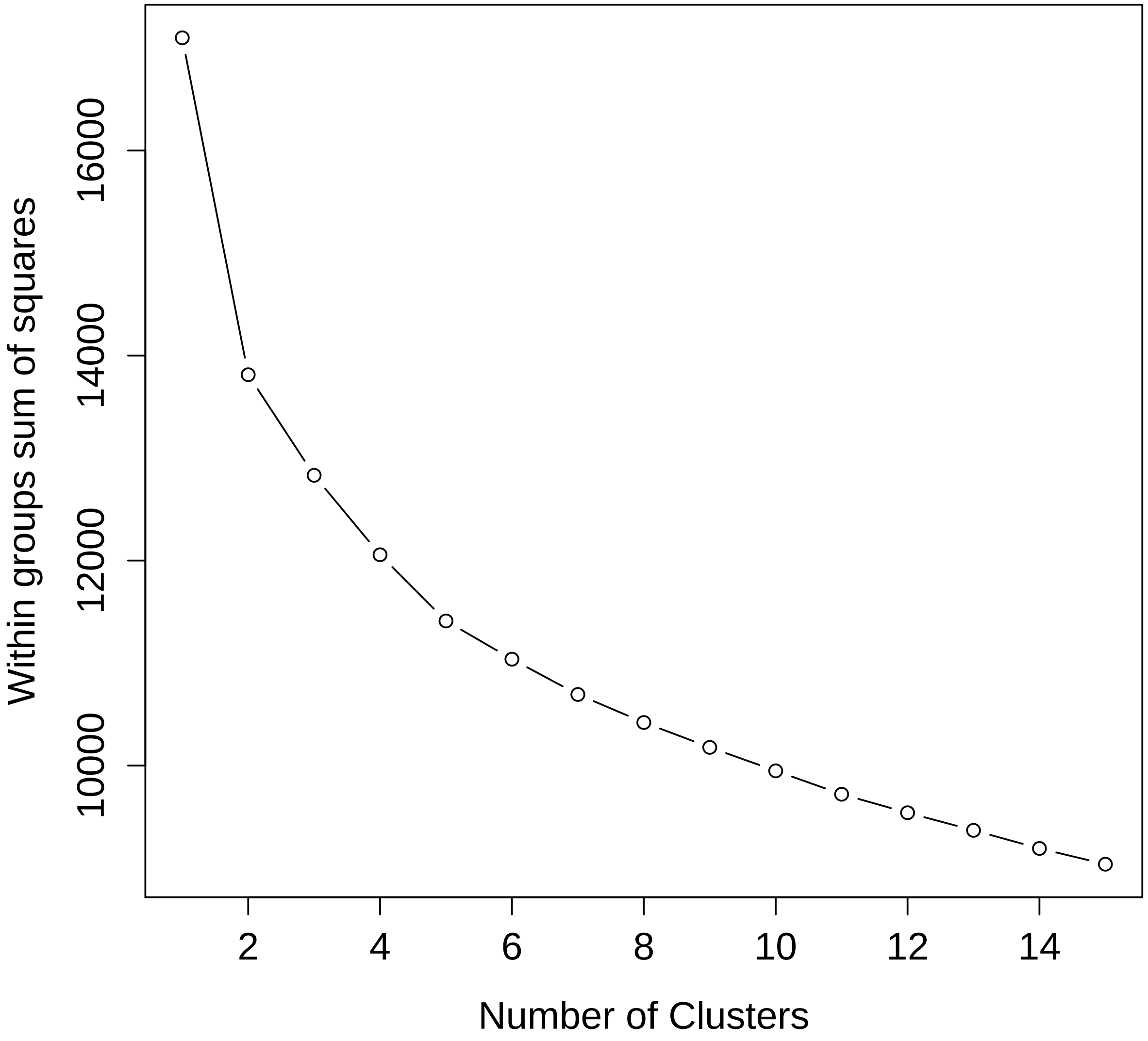

Figure 3: Scree test: WSS error by cluster size

Figure 3 shows no optimal elbow in scree test! Maybe 2-3?

2.1.2 Gap Statistic

Optimal cluster is the local maximum of the gap statistic, but it appears to be consistently increasing in Figure 4.

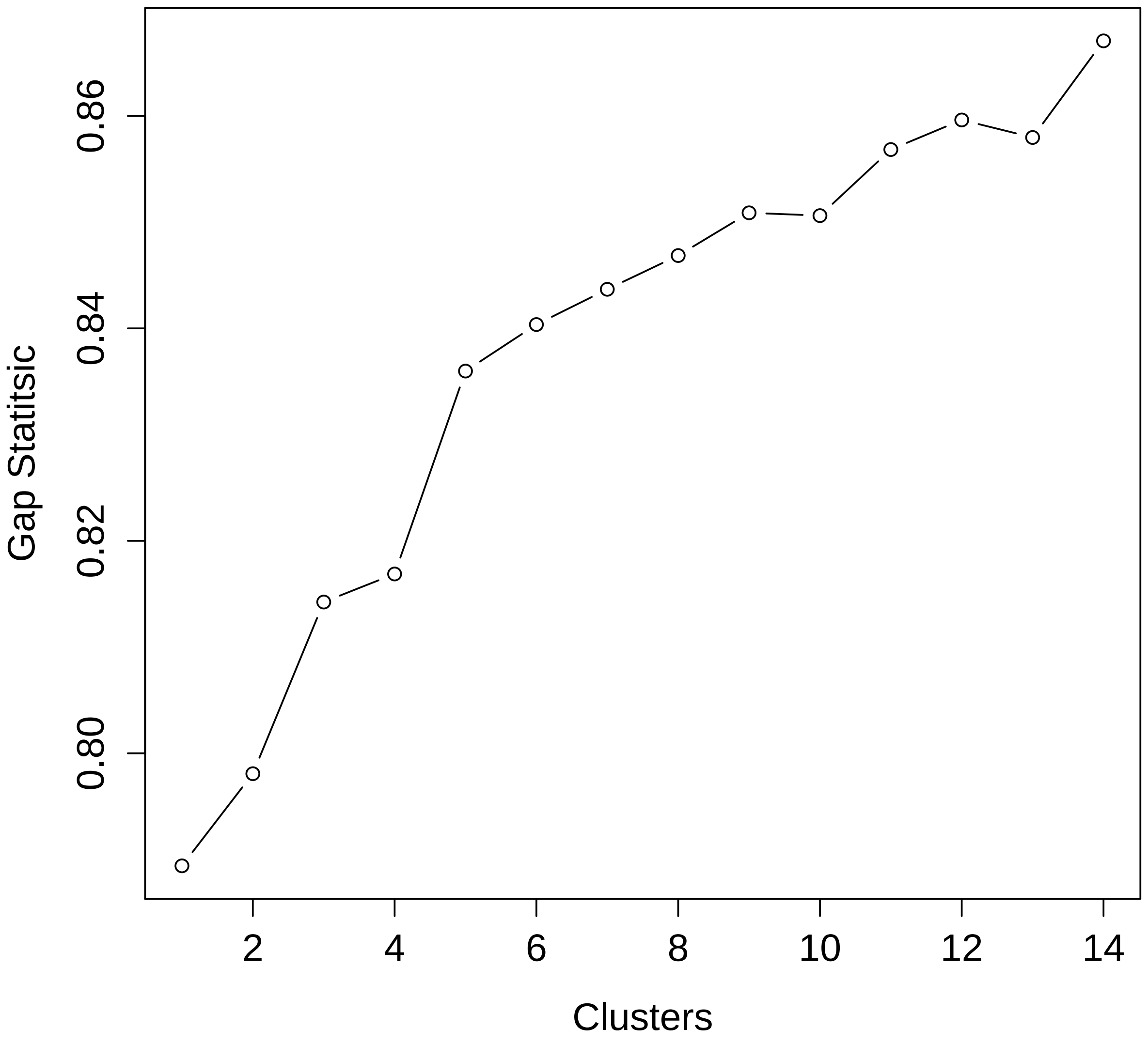

Figure 4: Gap statistic by cluster size

2.1.3 Average Silhouette Width

Figure 5 shows average silhouette width as being consistently under 0.25 for all clusters, implying the data is not well structured.

2.2 Cluster statistics

k	n	Within SS	sum(Within SS)
2	229/672	6118/7695	13813
3	333/134/434	4669/40009/4154	12832
4	79/394/275/153	2367/3357/3454/2880	12057

Table 3: Cluster statistics

2.3 Silhouette plots

Available in Figures 6, 7, and 8. Note: constructed with standardized z-score data.

2.4 Decision trees based on clusters

k	CP[3]	CV Xerror[4]	Root Feature	Root Error	Figure
2	0.0218	0.113	axial $\geq$ 4.5	0.254	Figure 9
3	0.0107	0.191	pigd $\geq$ 2.5	0.518	Figure 10
4	0.0100	0.255	pigd $<$ 2.5	0.563	Figure 11

Table 4: k-kmeans decision trees statistics

2.5 Interpretation of Clusters

2.5.1 Cluster summaries

Available in Figures 12, 13, and 14. Error bar is standard error.

2.5.2 Interpretation

$k = 2$ seems too basic. Cluster is organized solely by severity - all symptoms, including motor and nonmotor, are higher in severity in cluster 1, and lower in cluster 2. Quite consistently, groups in cluster 1 are generally of slightly higher age and pd duration.

$k = 3$ seems like a further development of $k = 2$, where clusters are simply organized by linearly increasing severity.

$k = 4$ is where it gets interesting.

[3] Complexity Parameter

[4] 10-fold cross validation

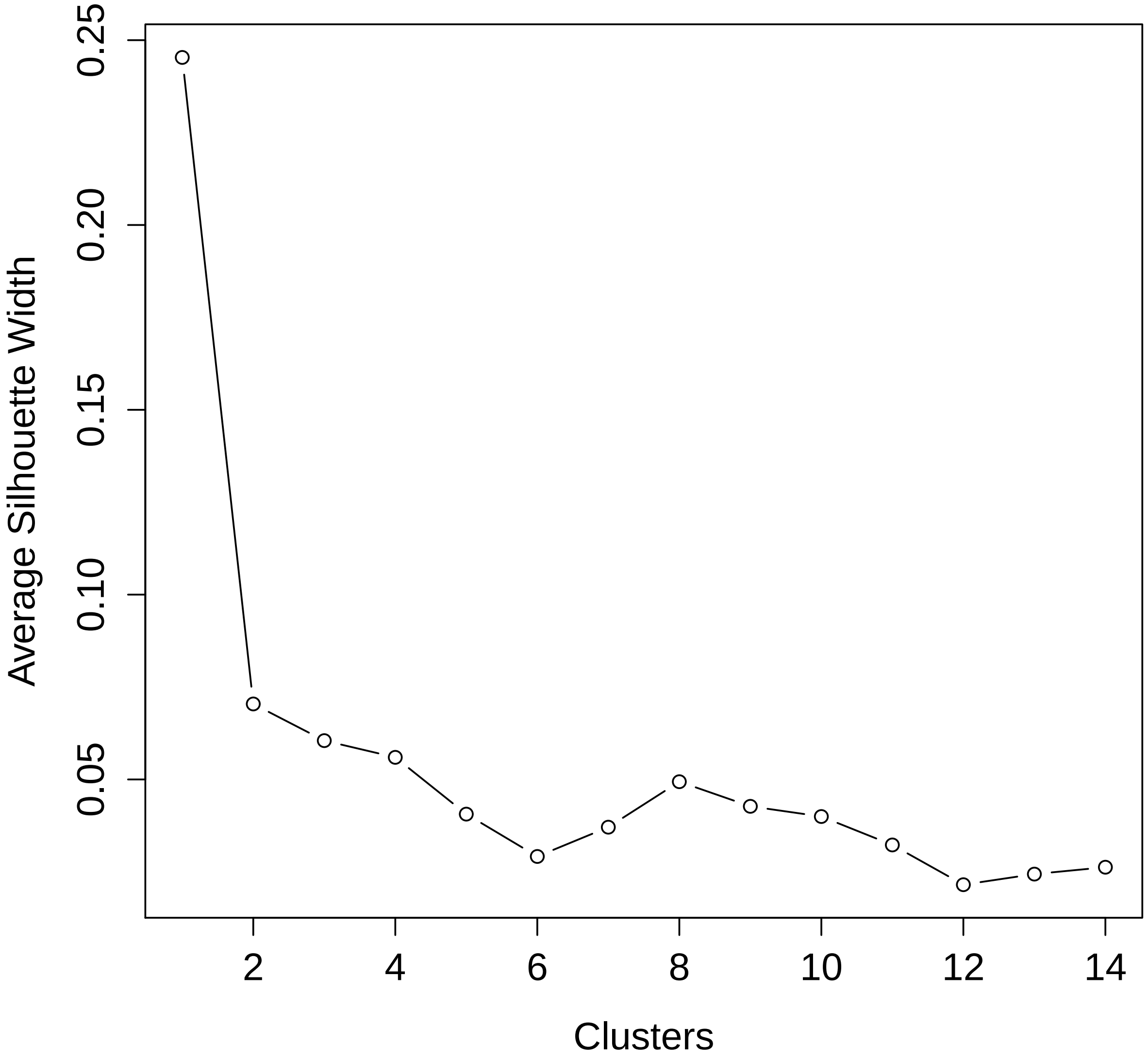

Figure 5: Average silhouette width by cluster size

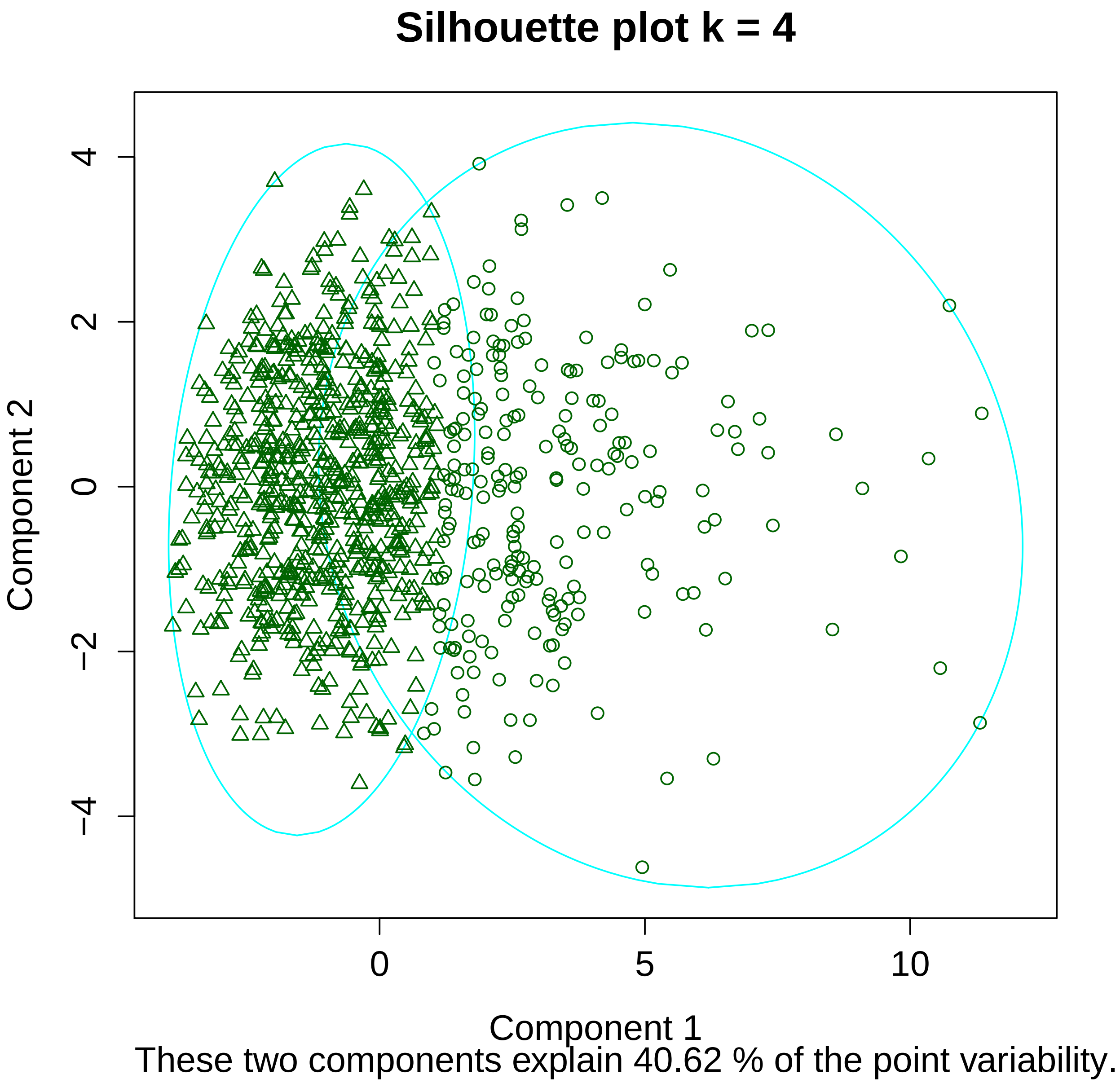

Figure 6: k-means cluster silhouette plot, $k = 2$

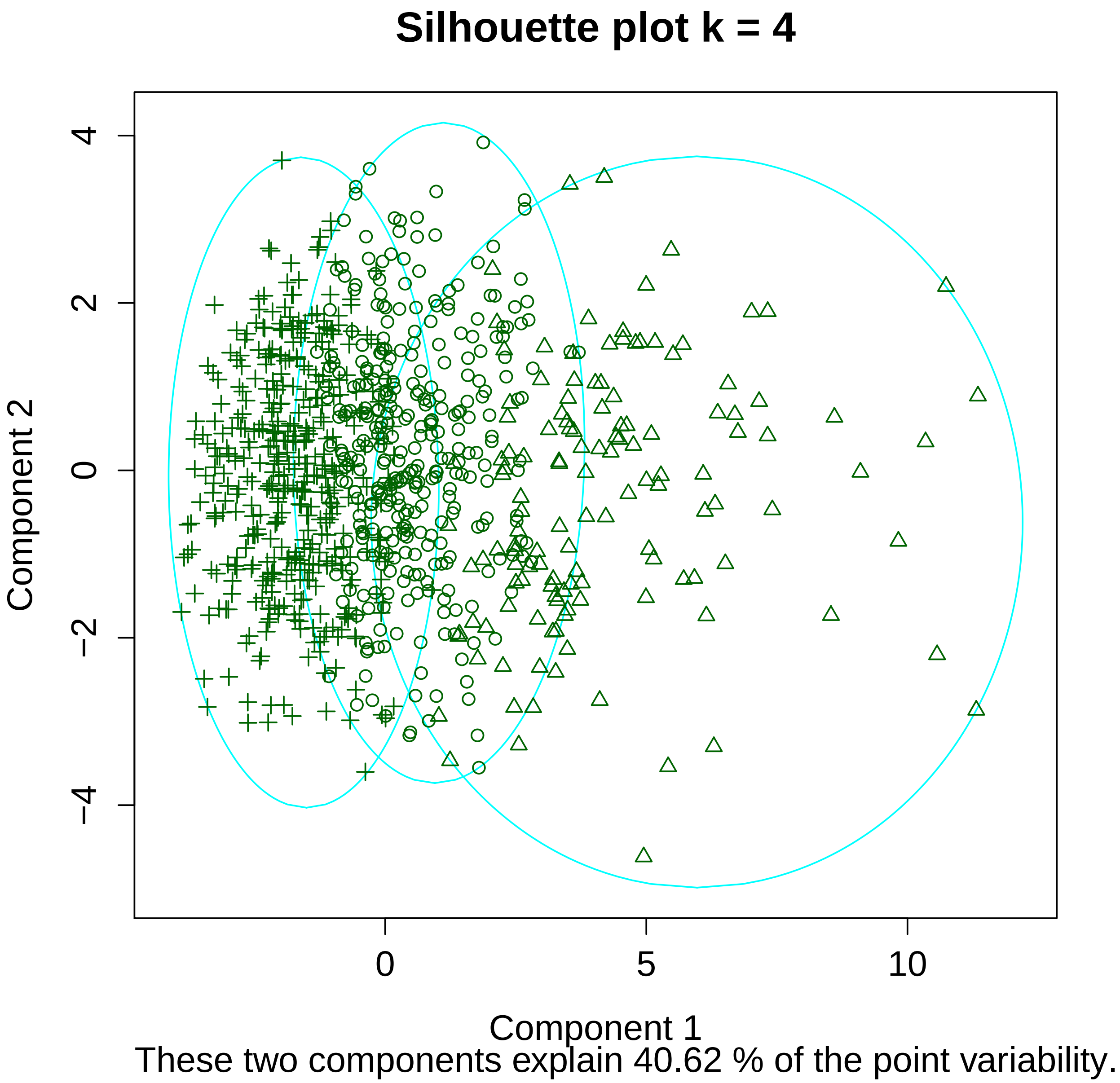

Figure 7: k-means cluster silhouette plot, $k = 3$

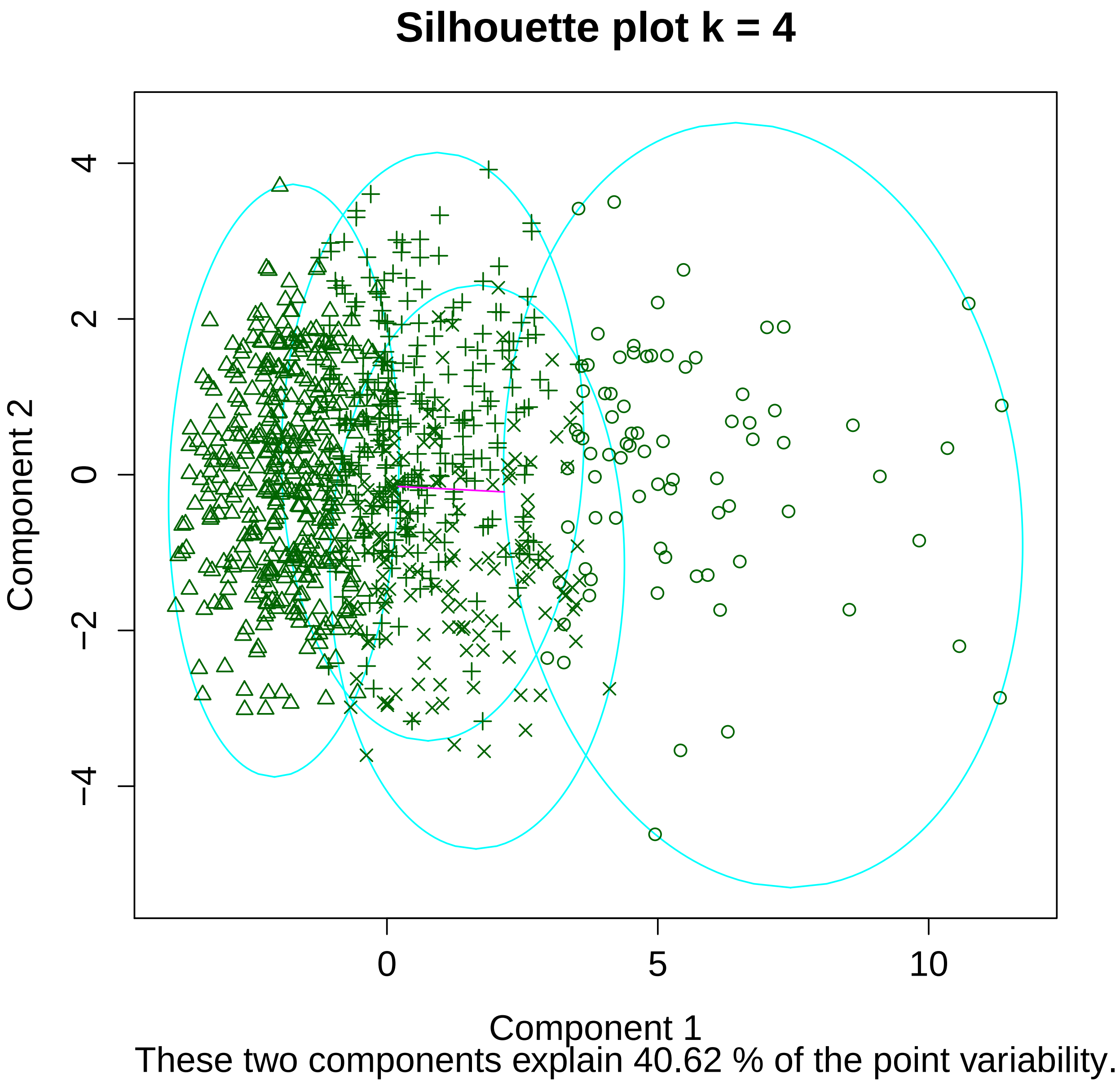

Figure 8: k-means cluster silhouette plot, $k = 4$

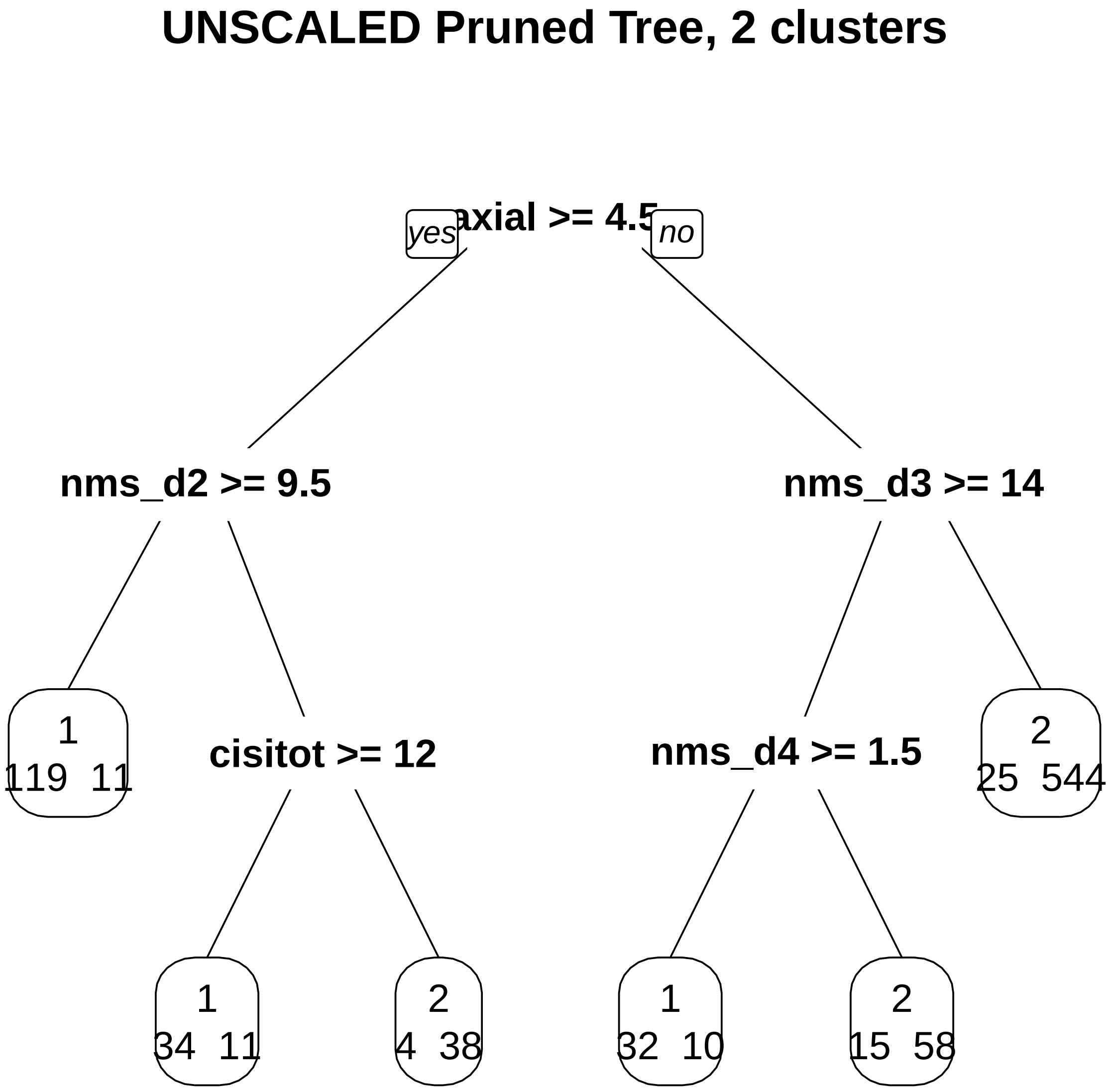

Figure 9: Decision Tree from k-means clustering, 2 clusters

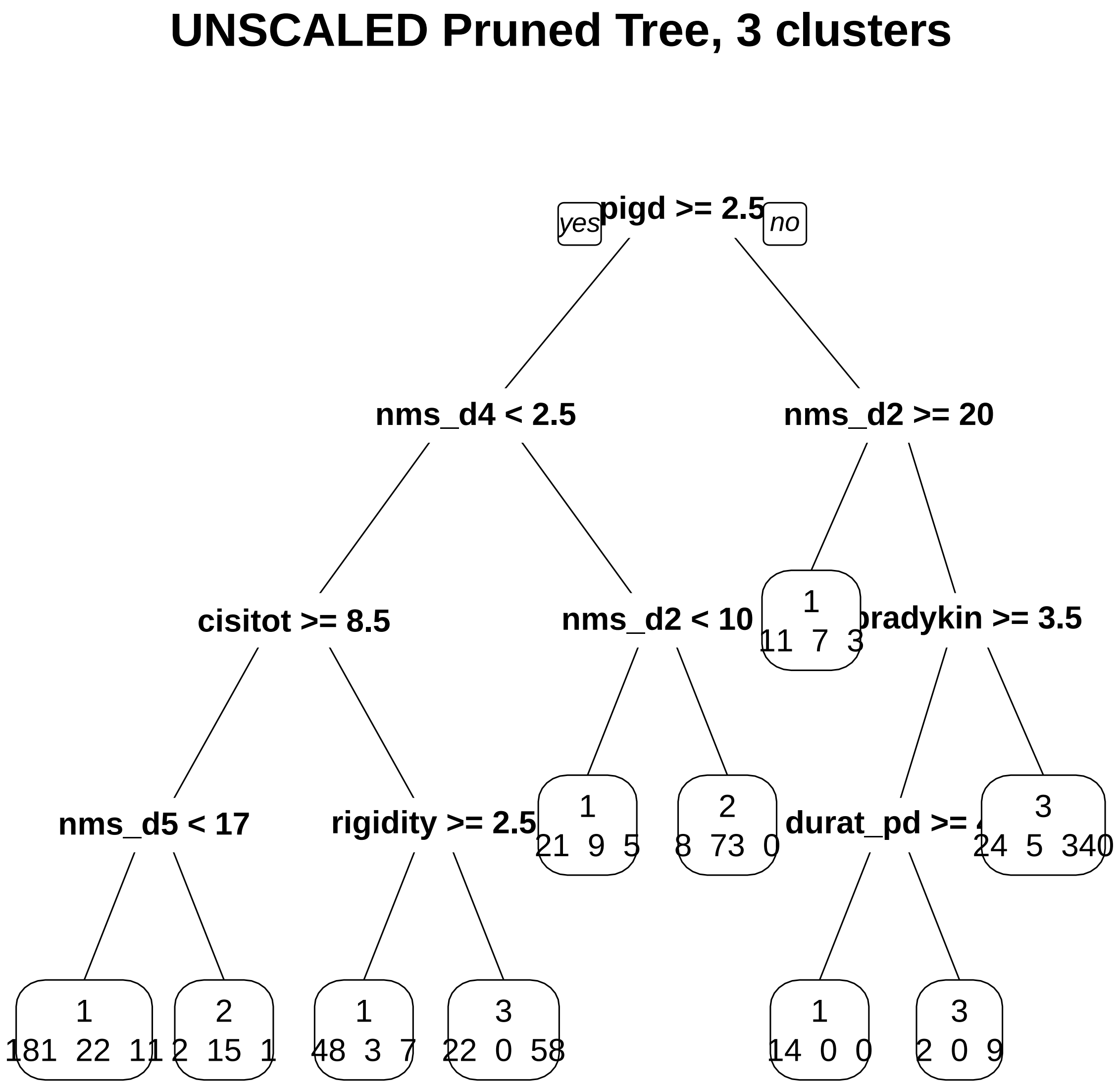

Figure 10: Decision Tree from k-means clustering, 3 clusters

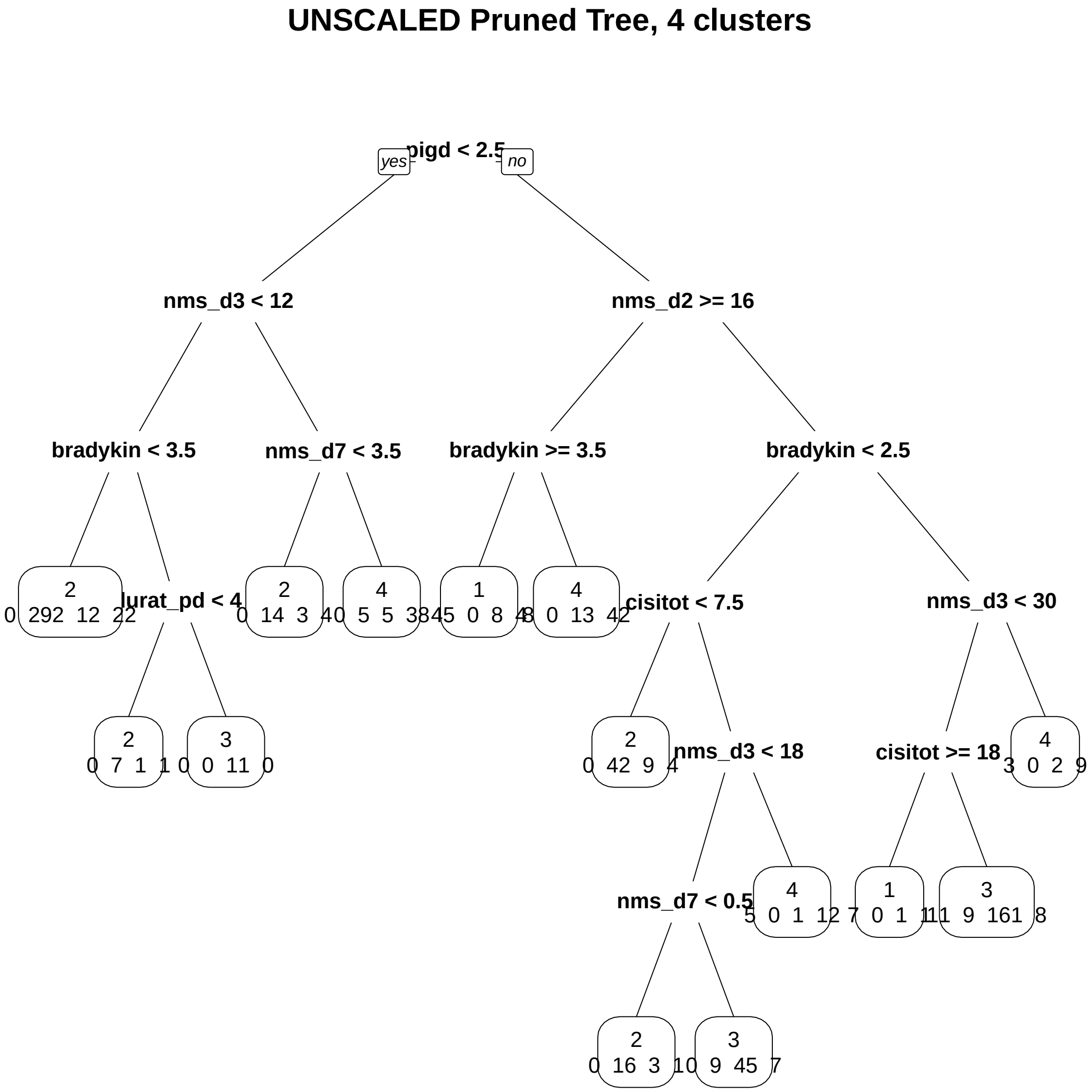

Figure 11: Decision Tree from k-means clustering, 4 clusters

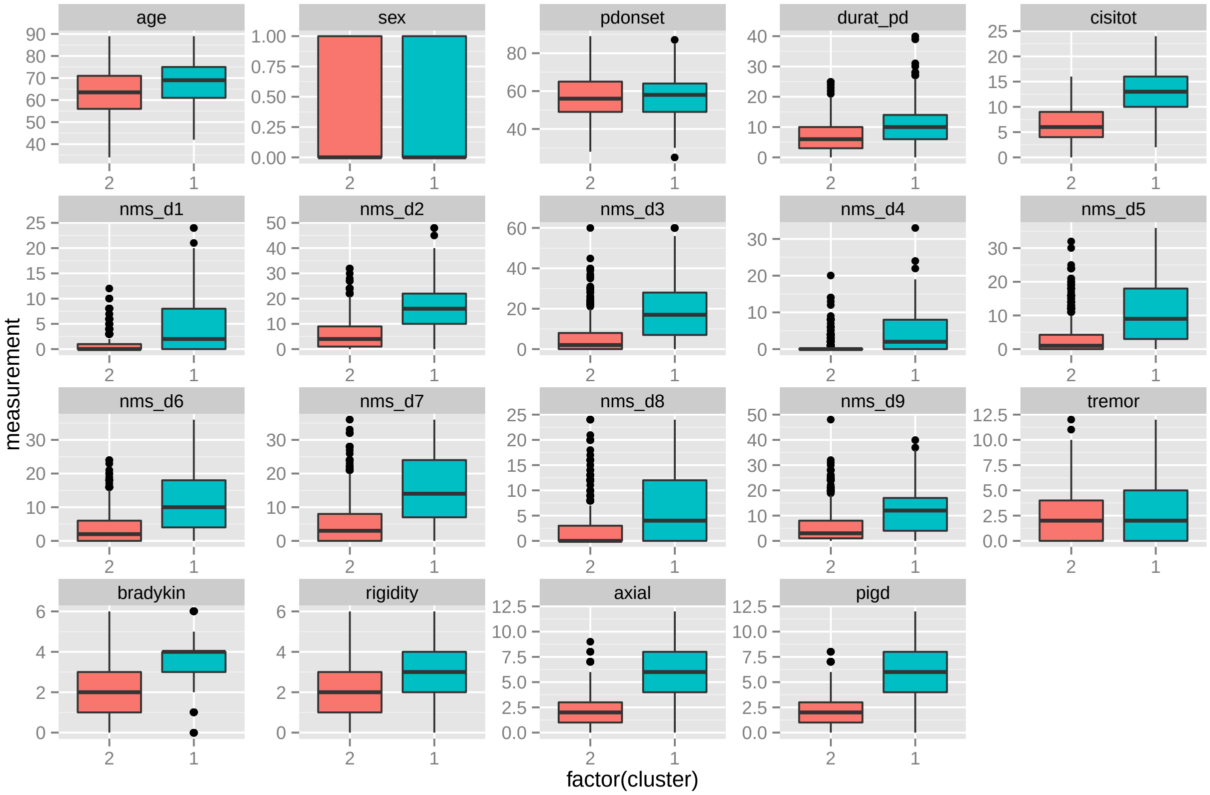

Figure 12: Cluster Summaries, $k = 2$

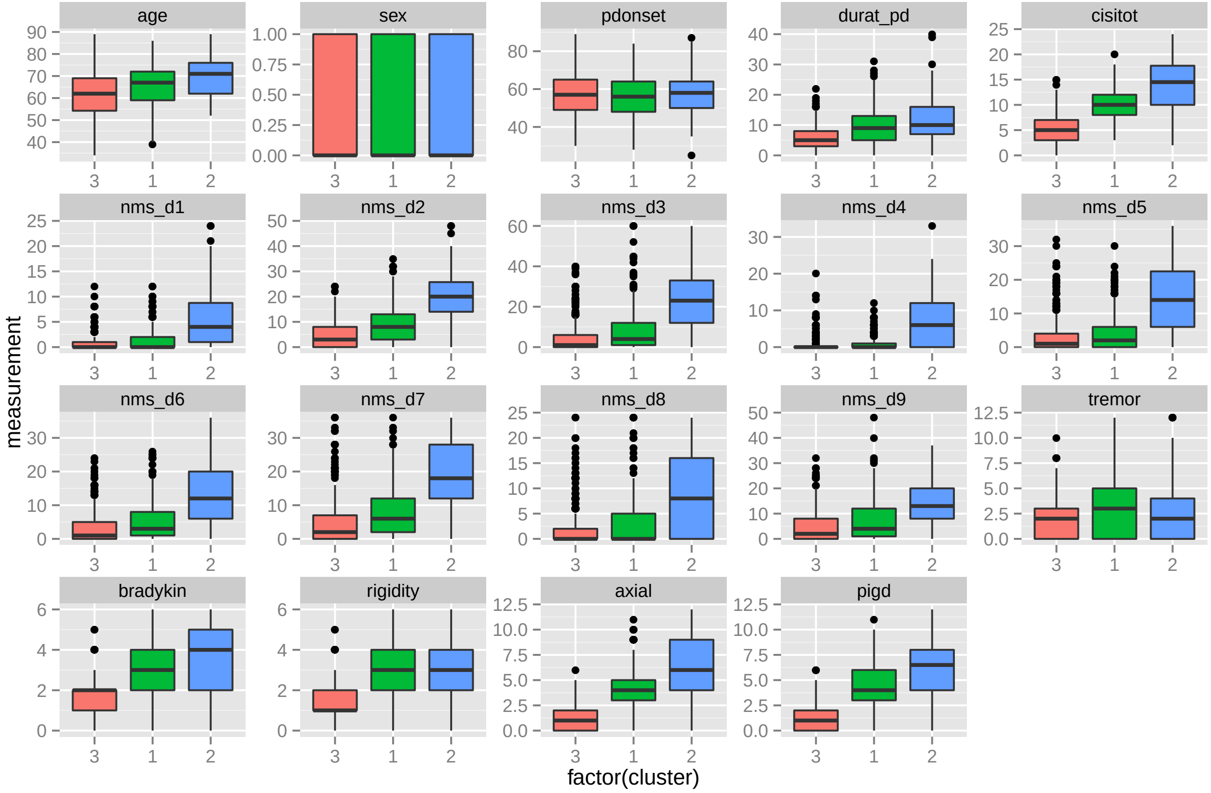

Figure 13: Cluster Summaries, $k = 3$

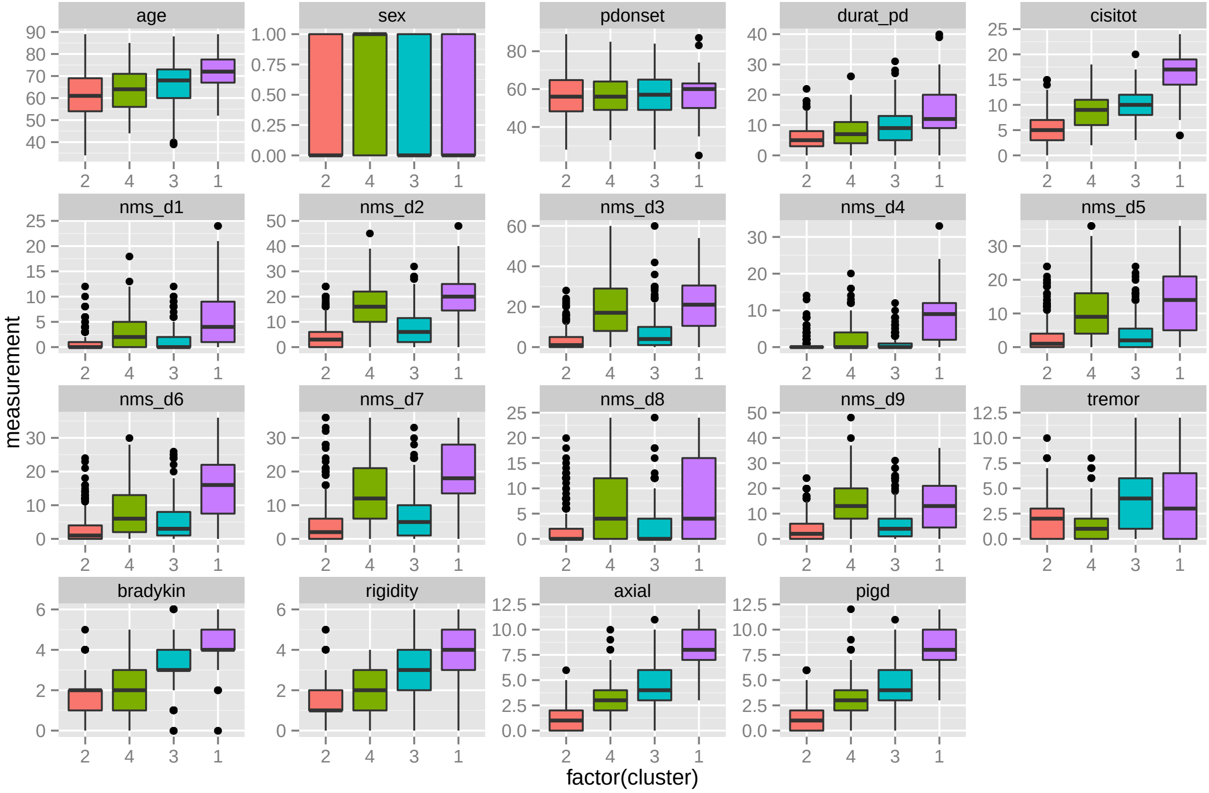

Figure 14: Cluster Summaries, $k = 4$

3 Affinity Propagation

3.1 Clustering

Package apcluster was used. Distance matrix was the negative euclidean squared distance ($r = 2$).

AP with input preferences minimized ($q = 0$) resulted in 8 clusters. With the standard median input preferences ($q = 0.5$), algorithm failed to converge with default parameters. Even setting damping factor to 0.98, maximum iterations to 10000, and convergence iterations to 1000 failed to converge. Might need to try a longer run.

However, given that input preferences control how many clusters are found, I don't think it's very useful to have some dozen clusters running around.

3.1.1 Silhouette Plots

Silhouette plot in Figure 15 looks pretty weak, really. Tons of overlap between the clusters.

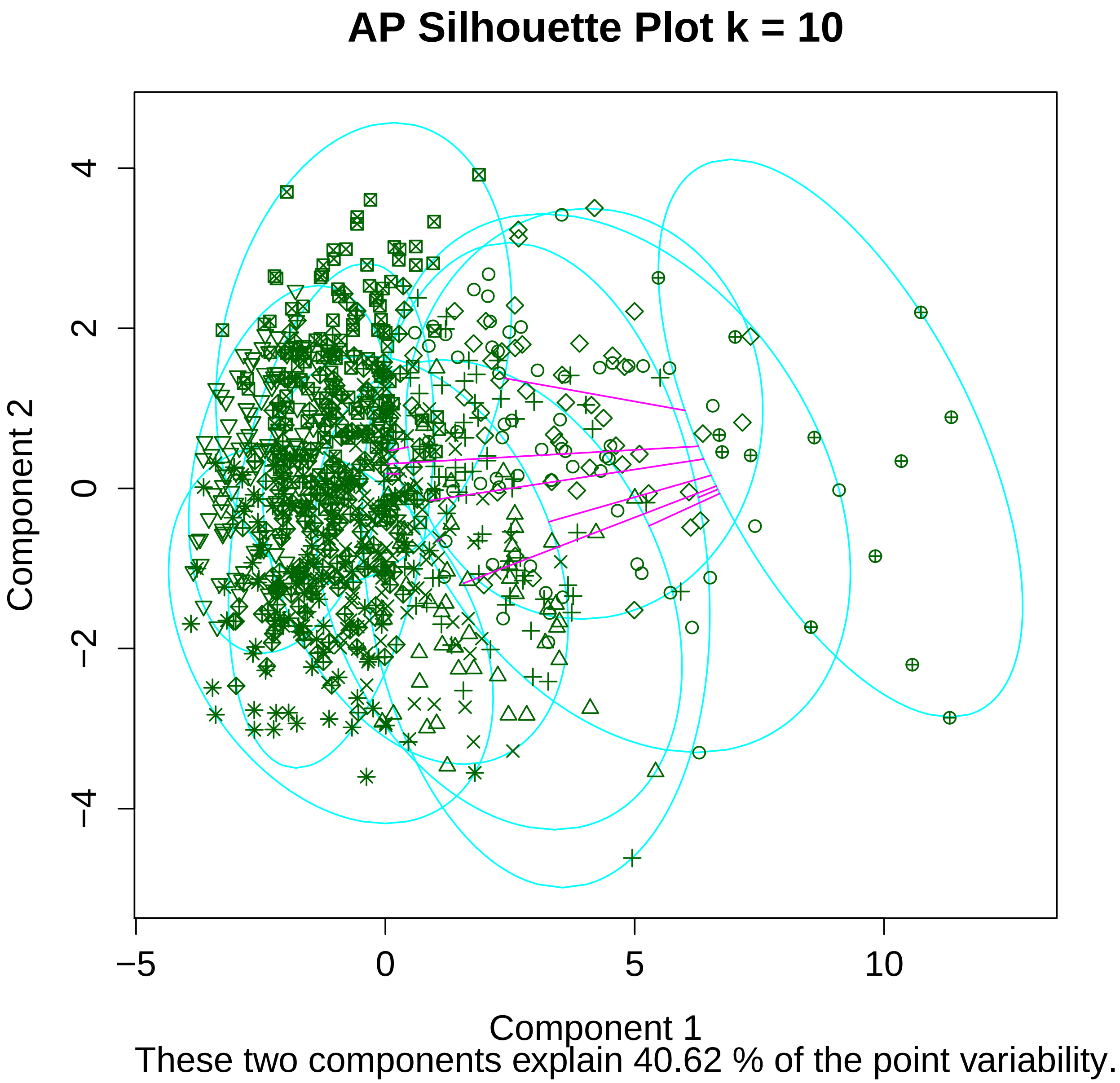

Figure 15: AP silhouette plot, $k = 8$

4 Hierarchical Clustering

4.1 Clustering

Four dissimilarity methods were used with a euclidean distance matrix. Dendrograms available in Figure 16

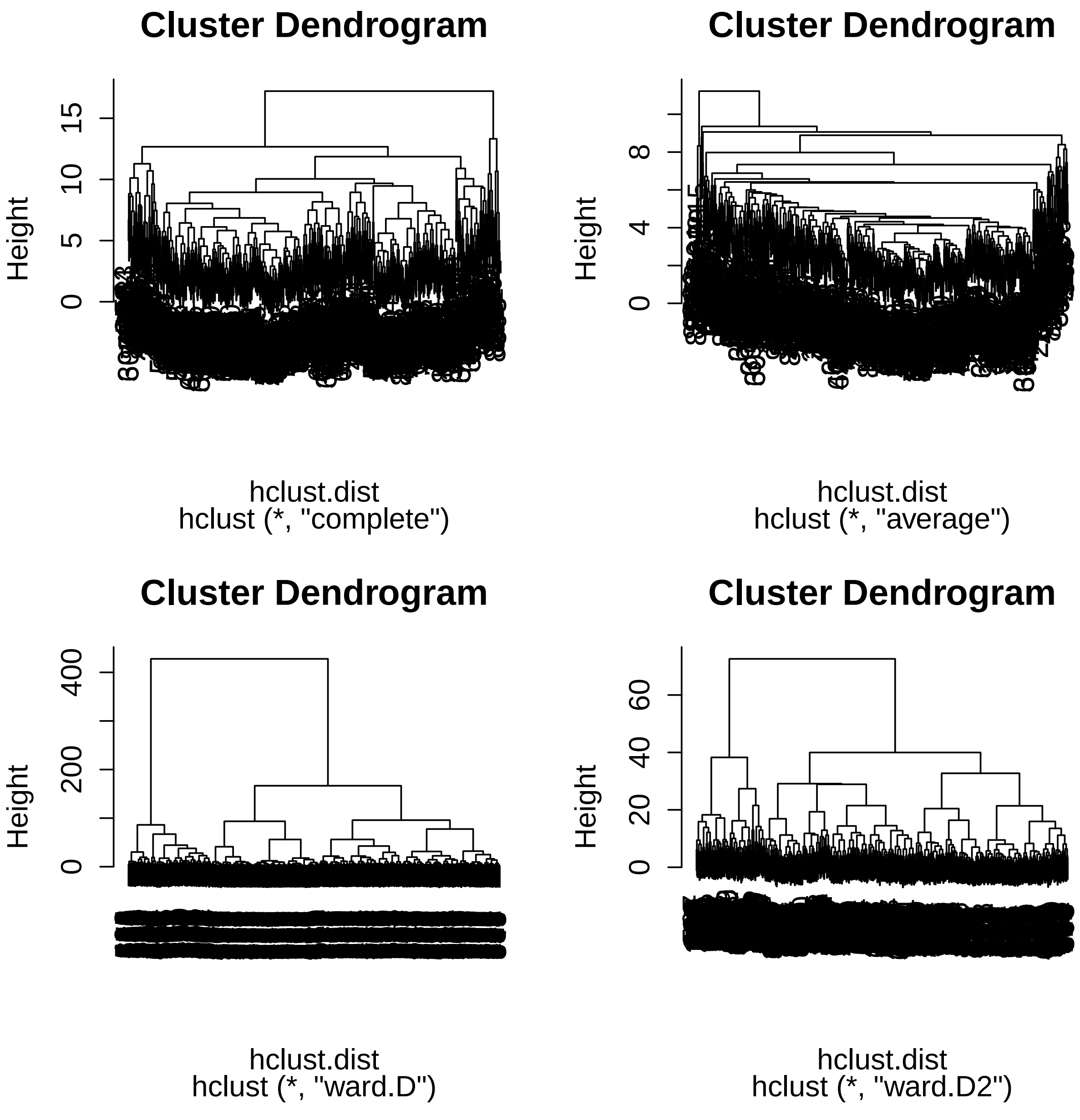

Figure 16: Dendrograms

Method	Condition	n	Figure
Complete	$k = 4$	4 (790/81/18/12)	17
Complete	`dynamicTreeCut`[5]	13 (255/99/77/64/62/58/56/46/44/41/37/32/30)	18
Ward	$k = 4$	4 (200/237/263/201)	19
Ward	$h = 100$	3 (437/263/201)	20

Table 5: Clusters from Tree Cutting

4.2 Cutting Trees

4.3 Interpretation

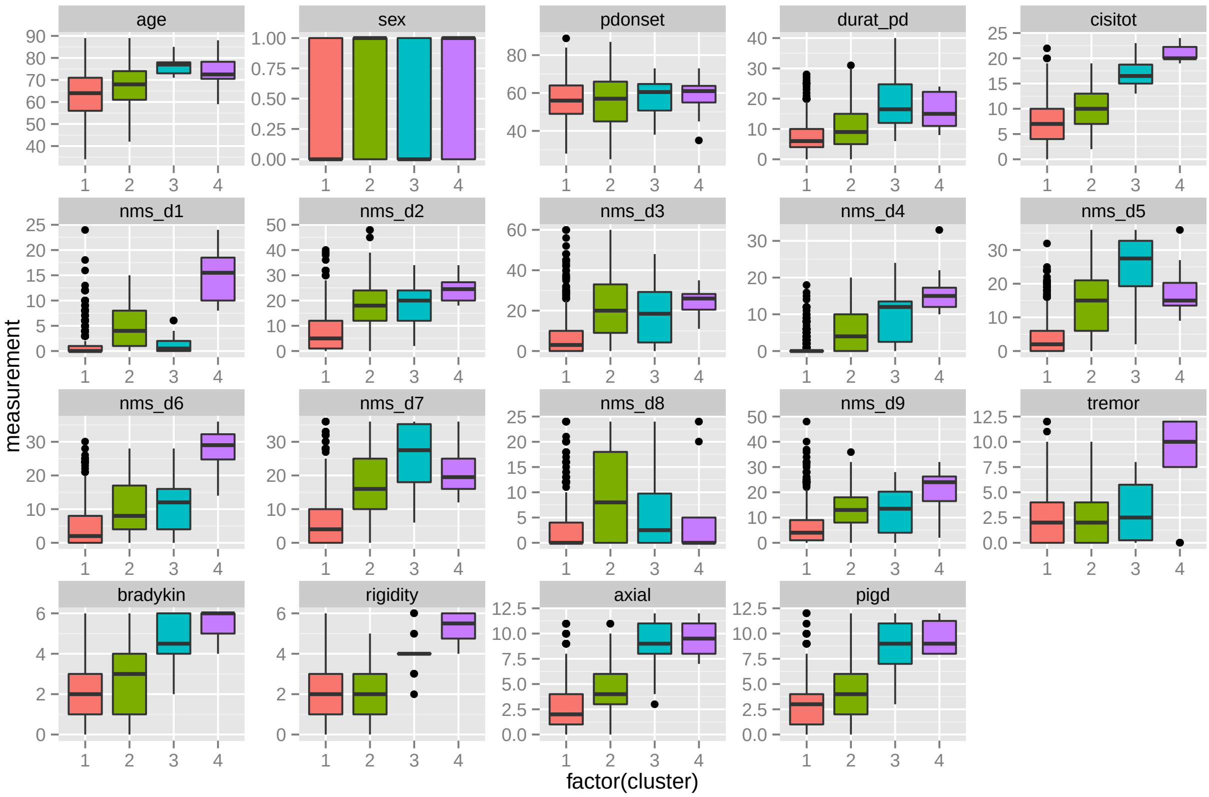

Figure 17: Using maximum (complete linkage) dissimilarity, cutting tree for $k = 4$

4.4 Interpretation

Cluster sizes are available in Table 6

Boxplot summary of clusters available in Figure 21. **Discussion forthcoming.**

[5] Package `dynamicTreeCut` in R (Langfelder P, Zhang B, Horvath S (2007)). Hybrid method, minimum cluster selection parameters

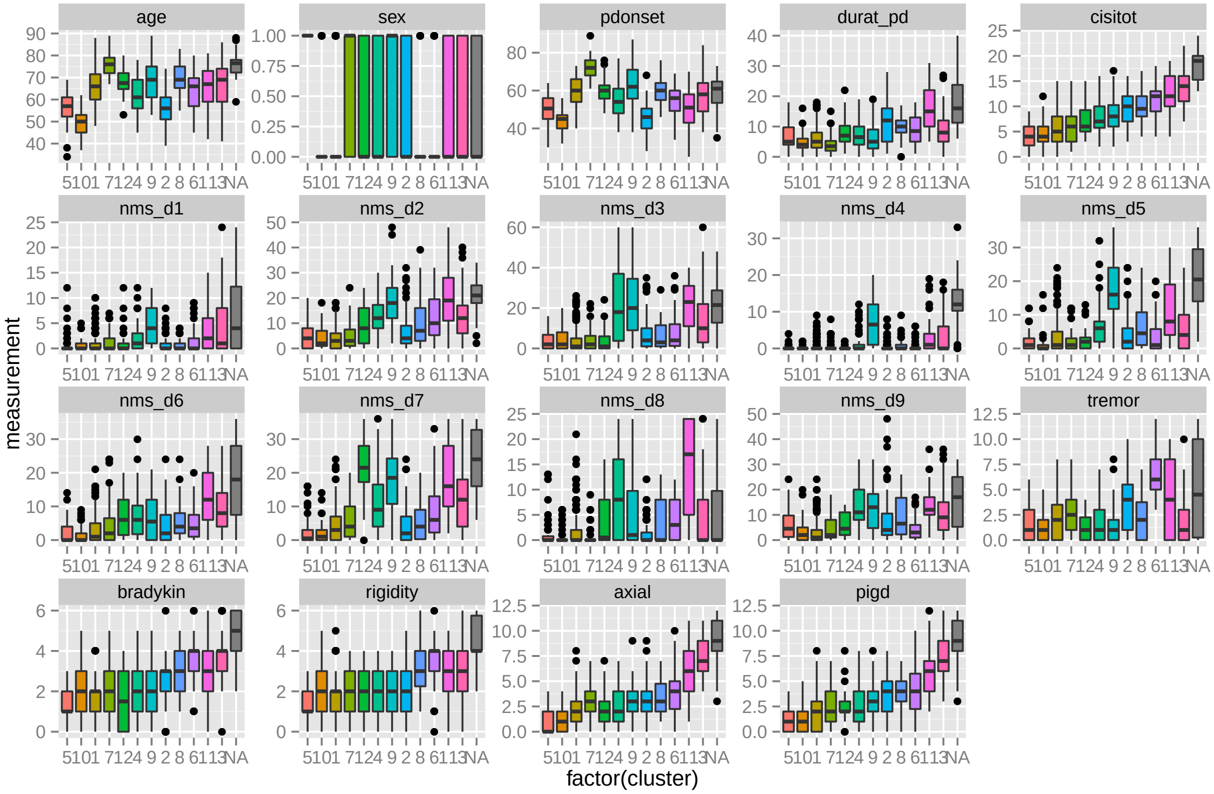

Figure 18: Using maximum (complete linkage) dissimilarity, cutting tree dynamically

5 Biclustering

Used BCBimax clustering algorithm. Clusters seem quite sparse.

6 Subspace clustering

7 Bayesian Networks

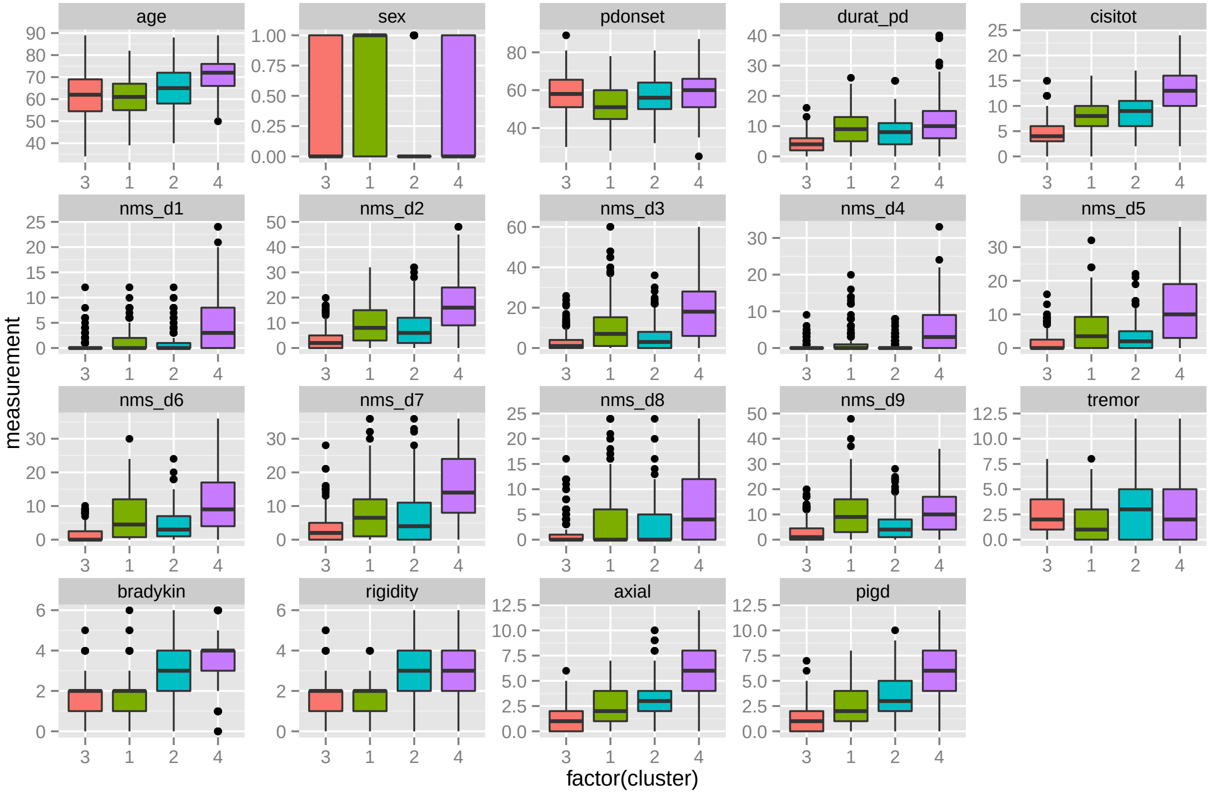

Figure 19: Using Ward (1963) dissimilarity, cutting tree for $k = 4$

Cluster	Size
1	63
2	53
3	85
4	122
5	48
6	126
7	123
8	102
9	166
10	13

Table 6: AP Cluster Sizes

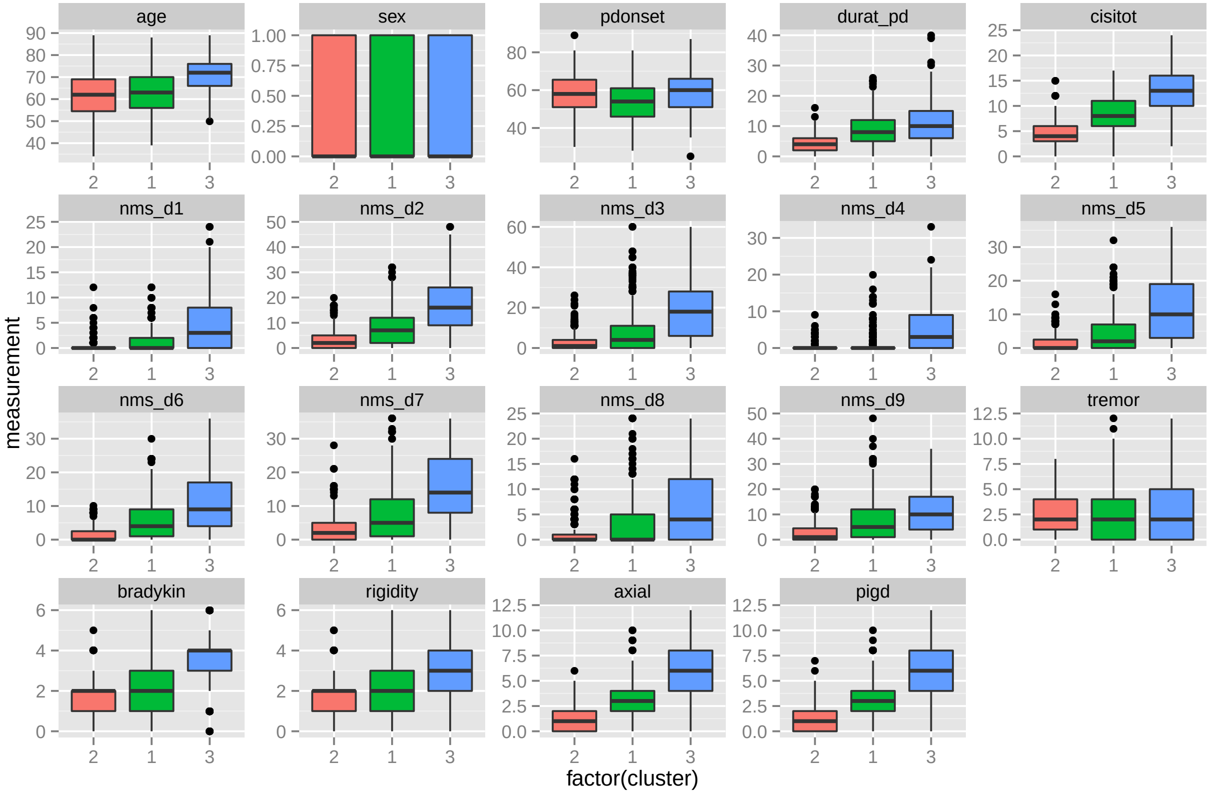

Figure 20: Using Ward (1963) dissimilarity, cutting tree at $h = 100$

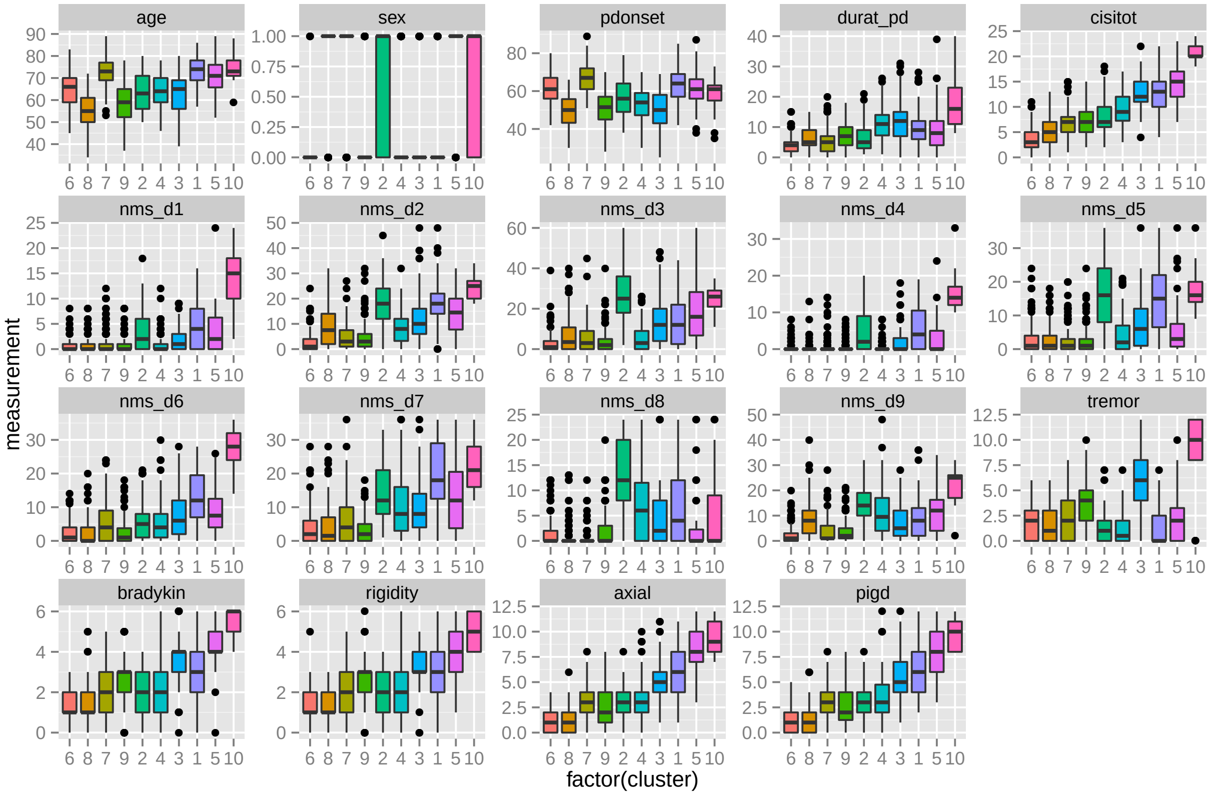

Figure 21: AP Boxplot Summaries

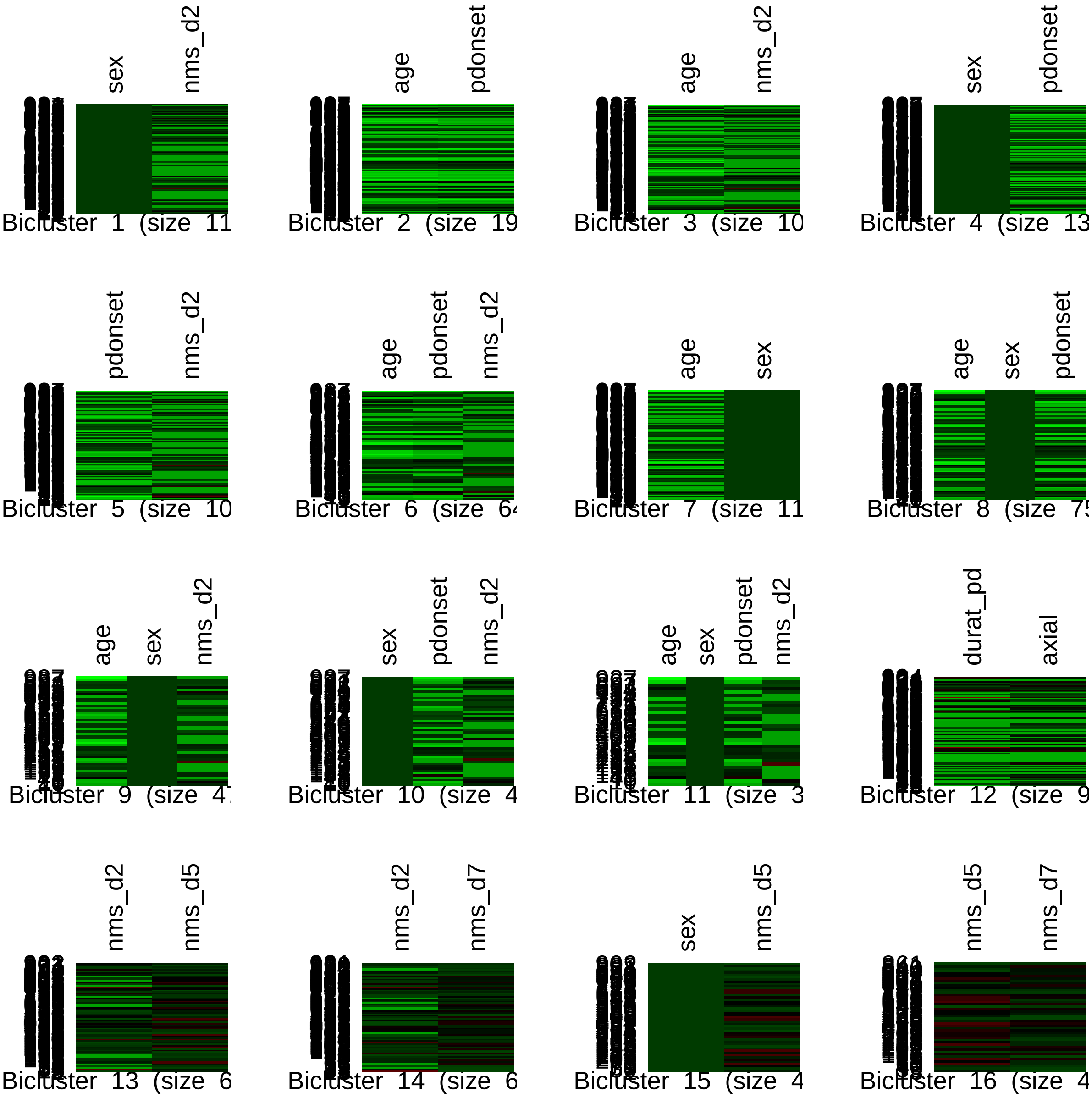

Figure 22: Biclustering heatmaps $N = 16$

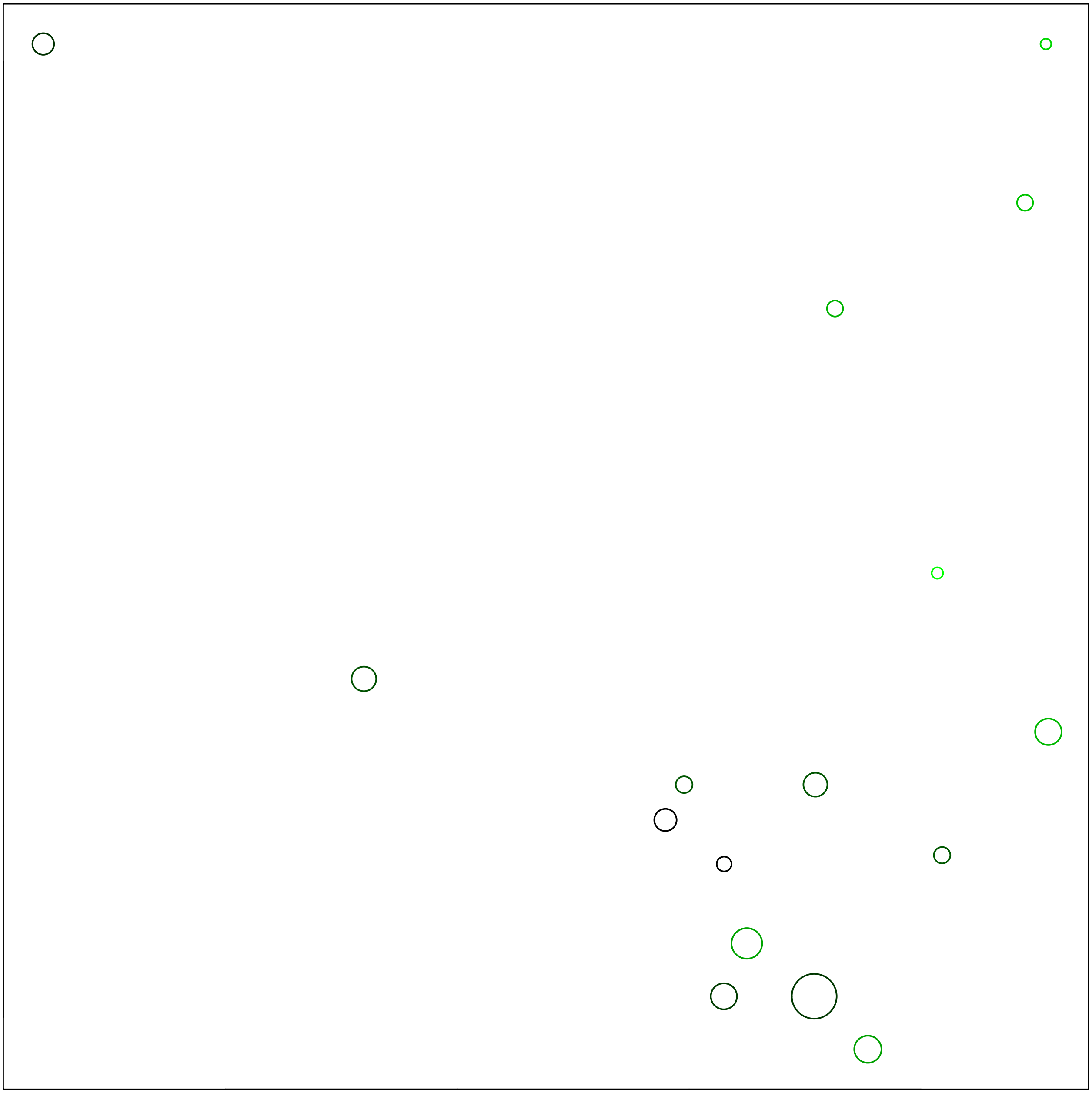

Figure 23: Bubbleplot $N = 16$