

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 Martín. 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

¹Impaired ability to adjust the body's position.

²Issues 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 $1 - 5 > 1$.

Individuals factor map (PCA) Variables factor map (PCA)

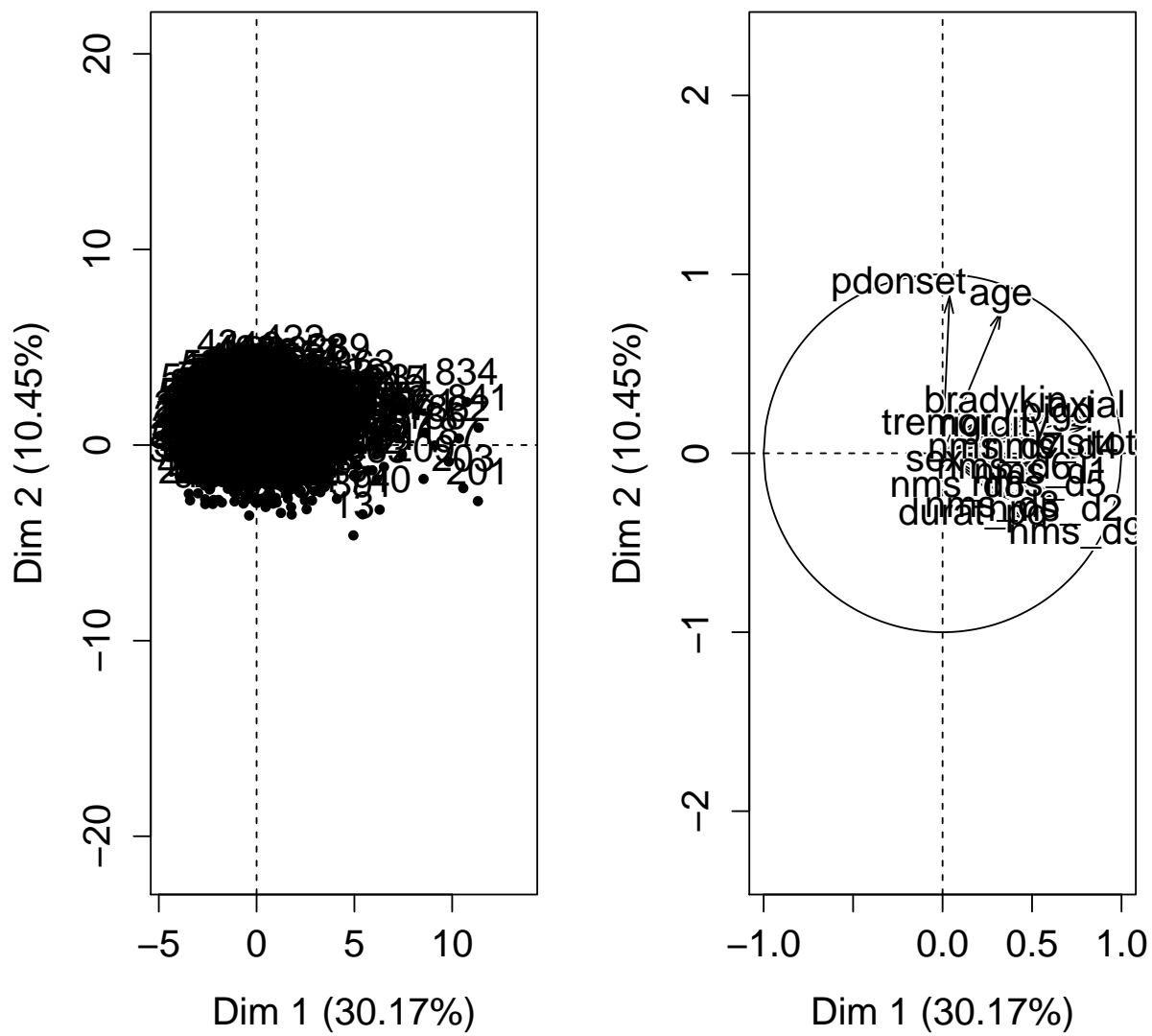


Figure 1: PCA Analysis

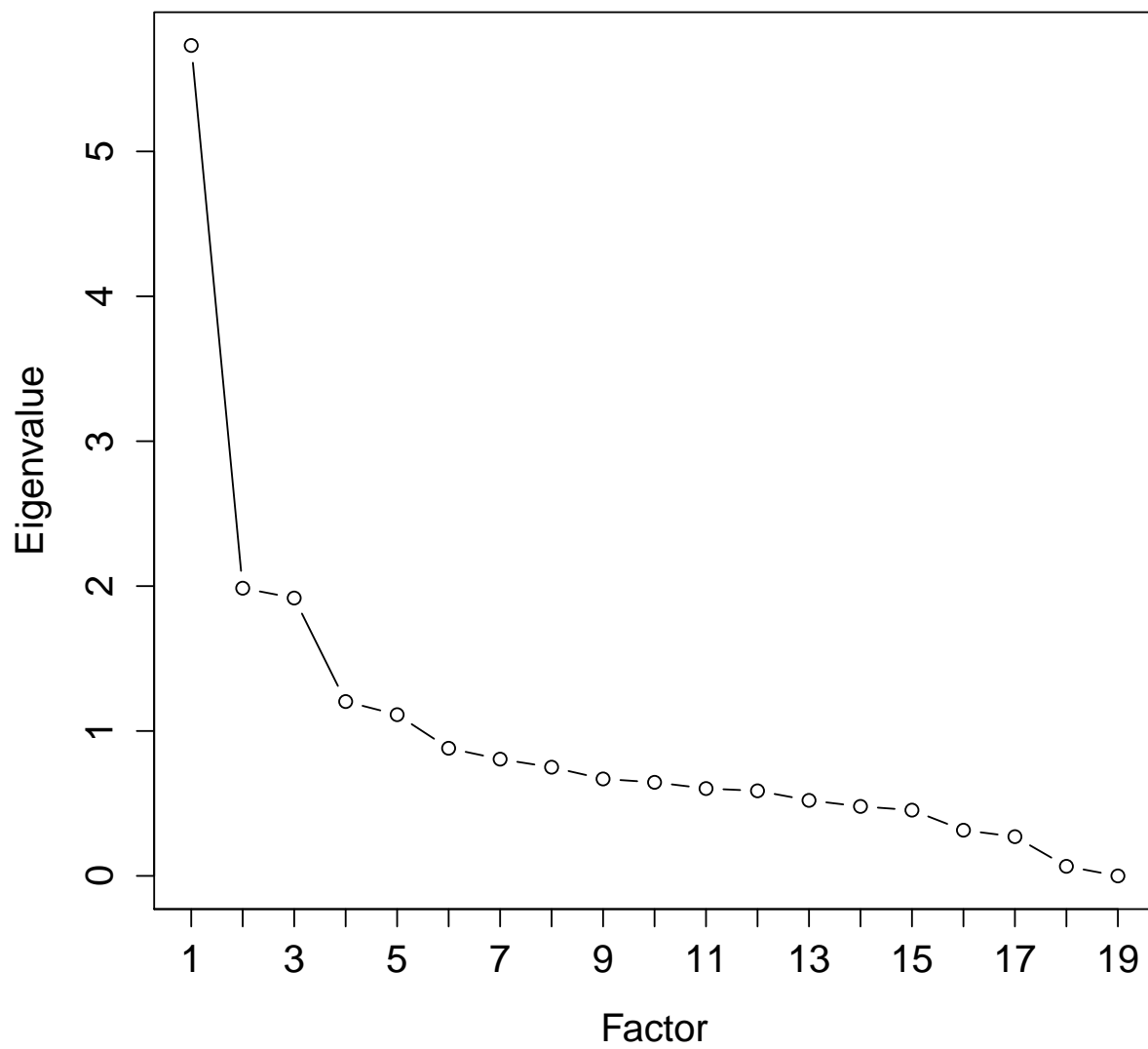


Figure 2: Scree test: eigenvalues by factor

2 k -means

2.1 Identifying optimal number of clusters

2.1.1 WSS Error Scree Test

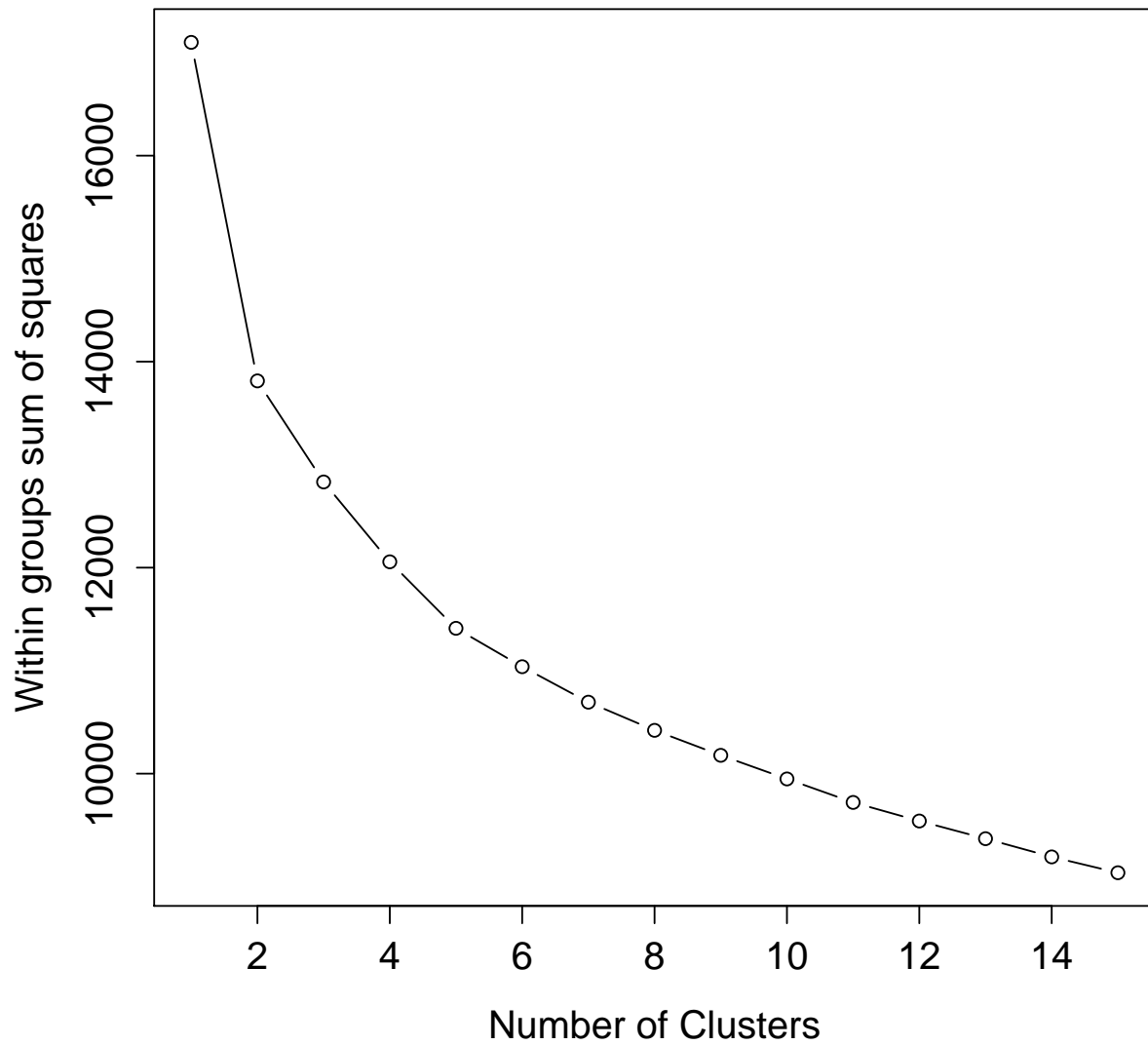


Figure 3: Scree test: WSS error by cluster size

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.

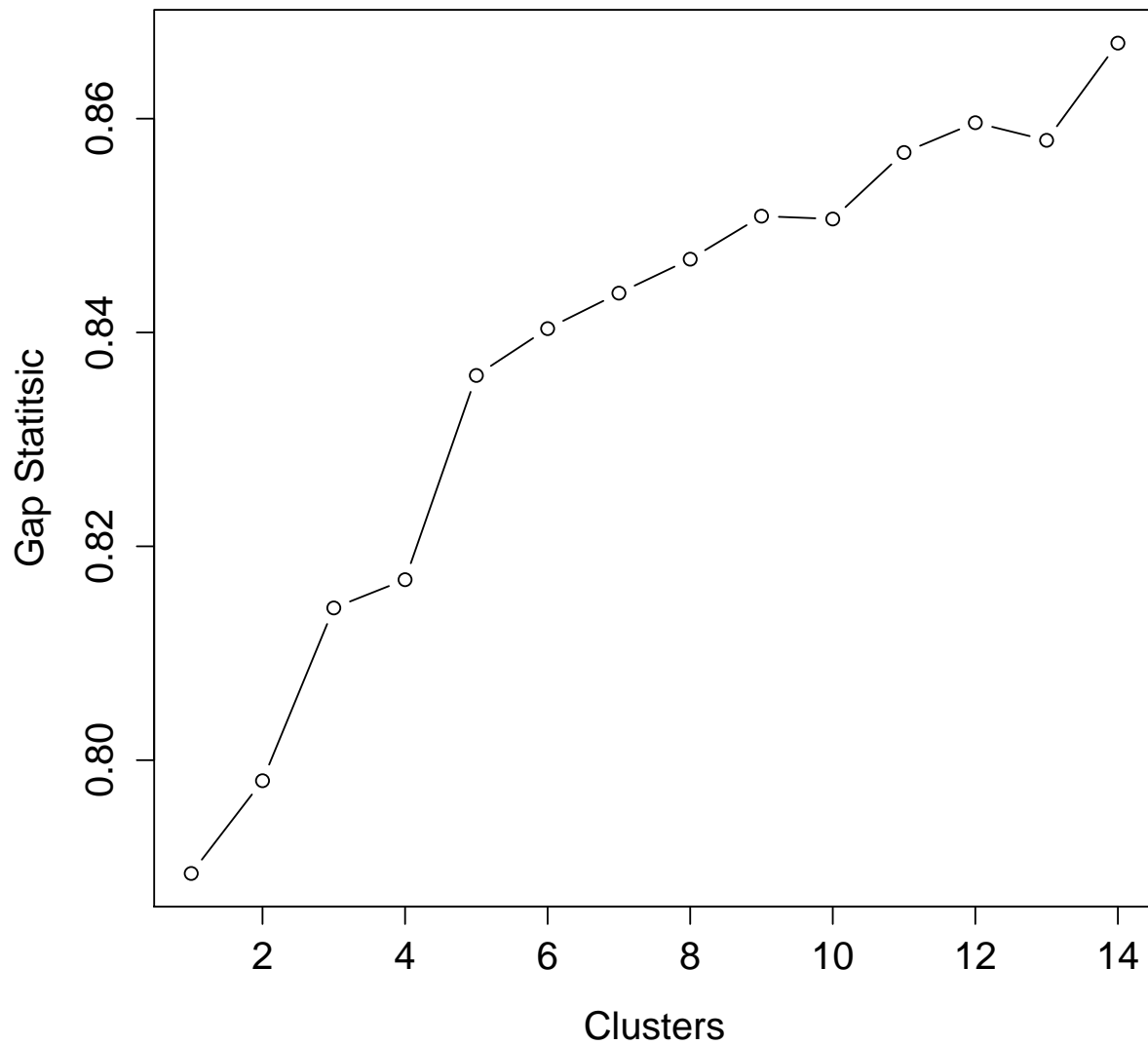


Figure 4: Gap statistic by cluster size

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 ≥ 4.5	0.254	Figure 9
3	0.0107	0.191	pigd ≥ 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.

2.5.3 Statistical Significance Tests, $k = 4$

Using one-way ANOVA for multiple means, we reject the null hypothesis that the means are the same with $p < 0.05$ for every variable *except* `pdonset`.

³Complexity Parameter

⁴10-fold cross validation

2.5.4 Ranked Features

Table 5: Features ranked by information gain

variable	information gain
axial	0.20640691
cisitot	0.20008571
pigd	0.18193982
nms_d2	0.13178572
nms_d9	0.12116024
bradykin	0.11966097
nms_d3	0.09421859
rigidity	0.09260628
nms_d5	0.07579997
nms_d4	0.07438784
nms_d6	0.06620599
nms_d7	0.05574956
nms_d1	0.05509838
tremor	0.04140473
nms_d8	0.03786173
durat_pd	0.02794420
age	0.00000000
sex	0.00000000
pdonset	0.00000000

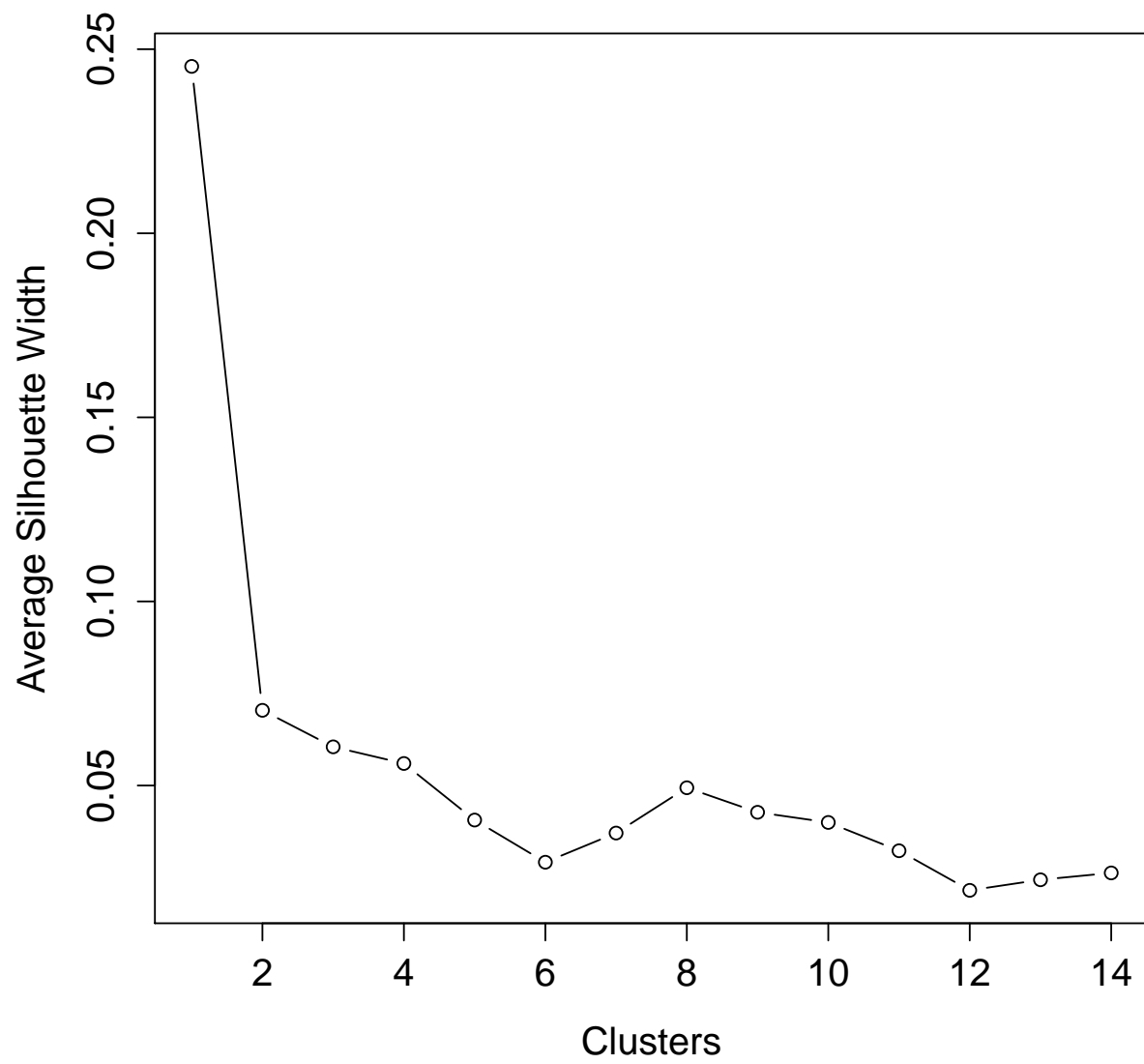


Figure 5: Average silhouette width by cluster size

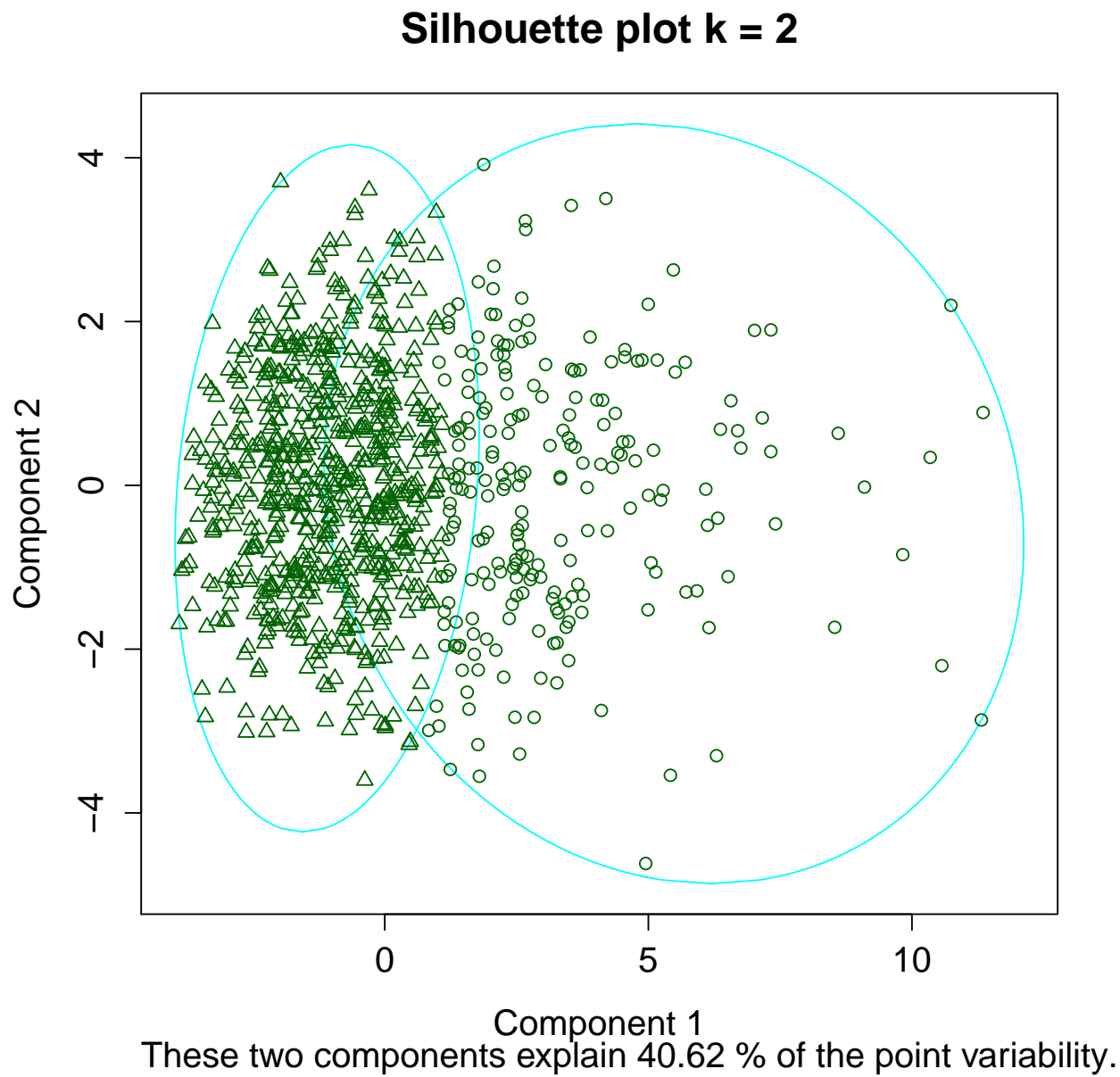


Figure 6: k -means cluster silhouette plot, $k = 2$

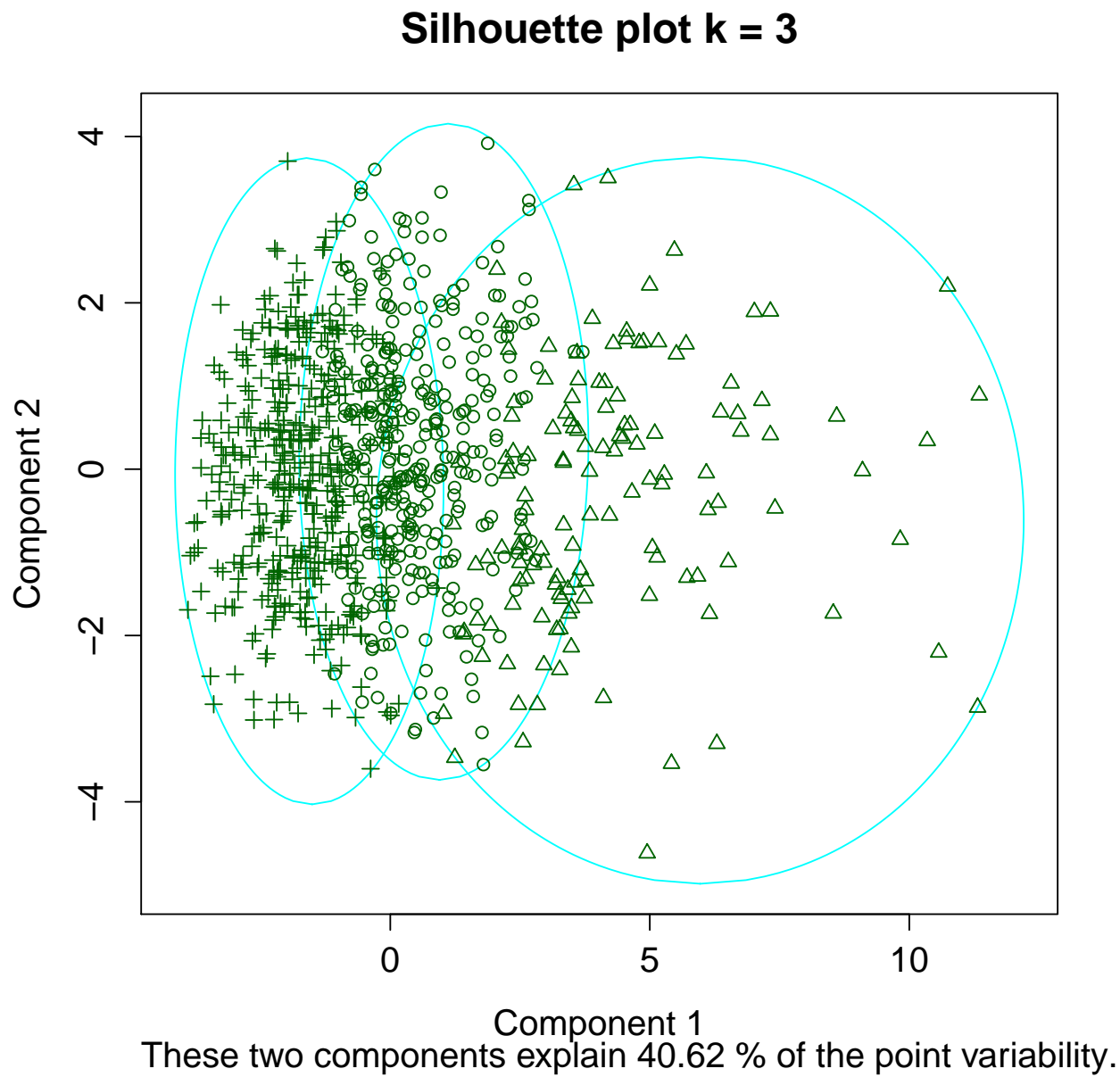
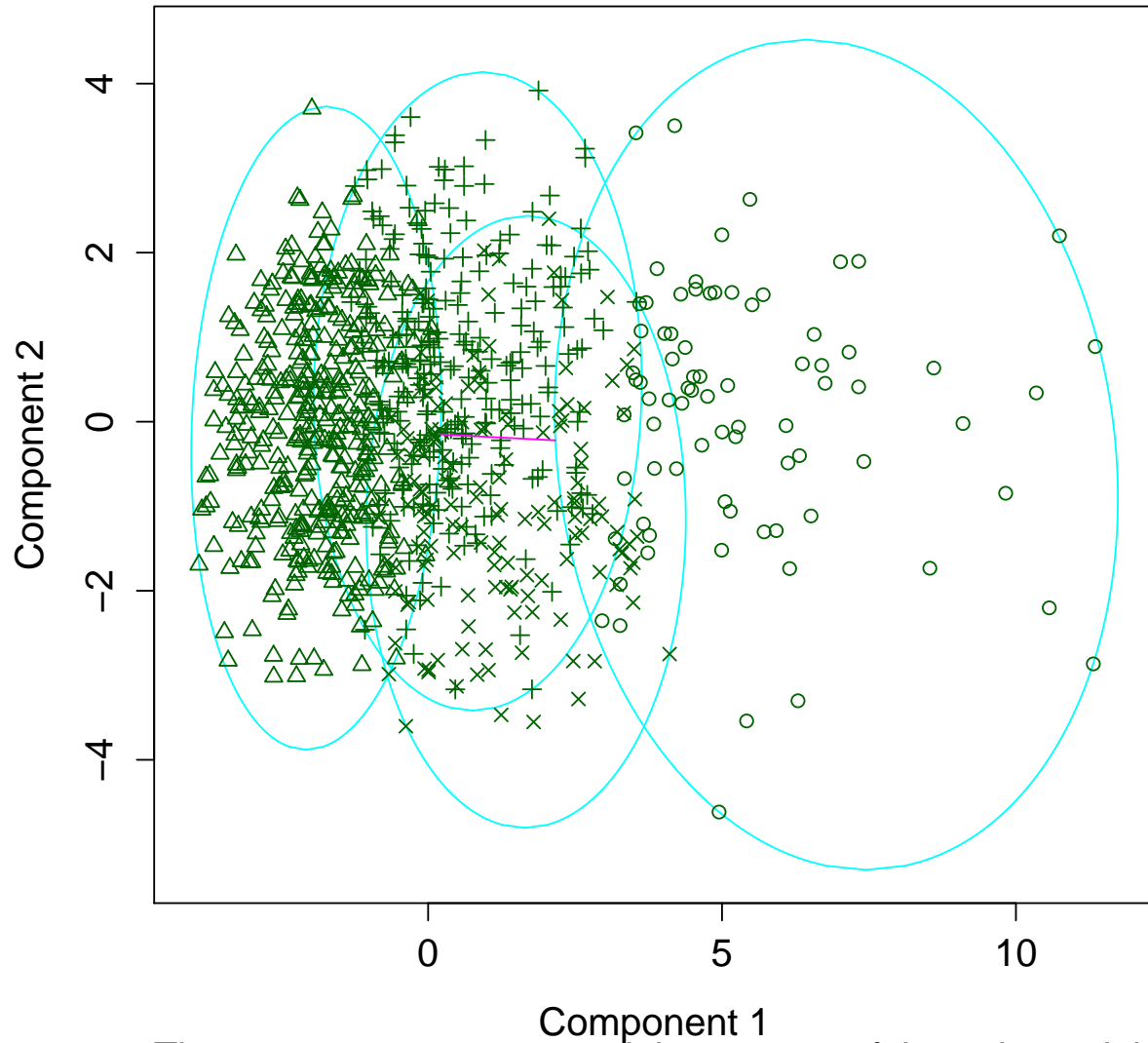


Figure 7: k -means cluster silhouette plot, $k = 3$

Silhouette plot $k = 4$



These two components explain 40.62 % of the point variability.

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

UNSCALED Pruned Tree, 2 clusters

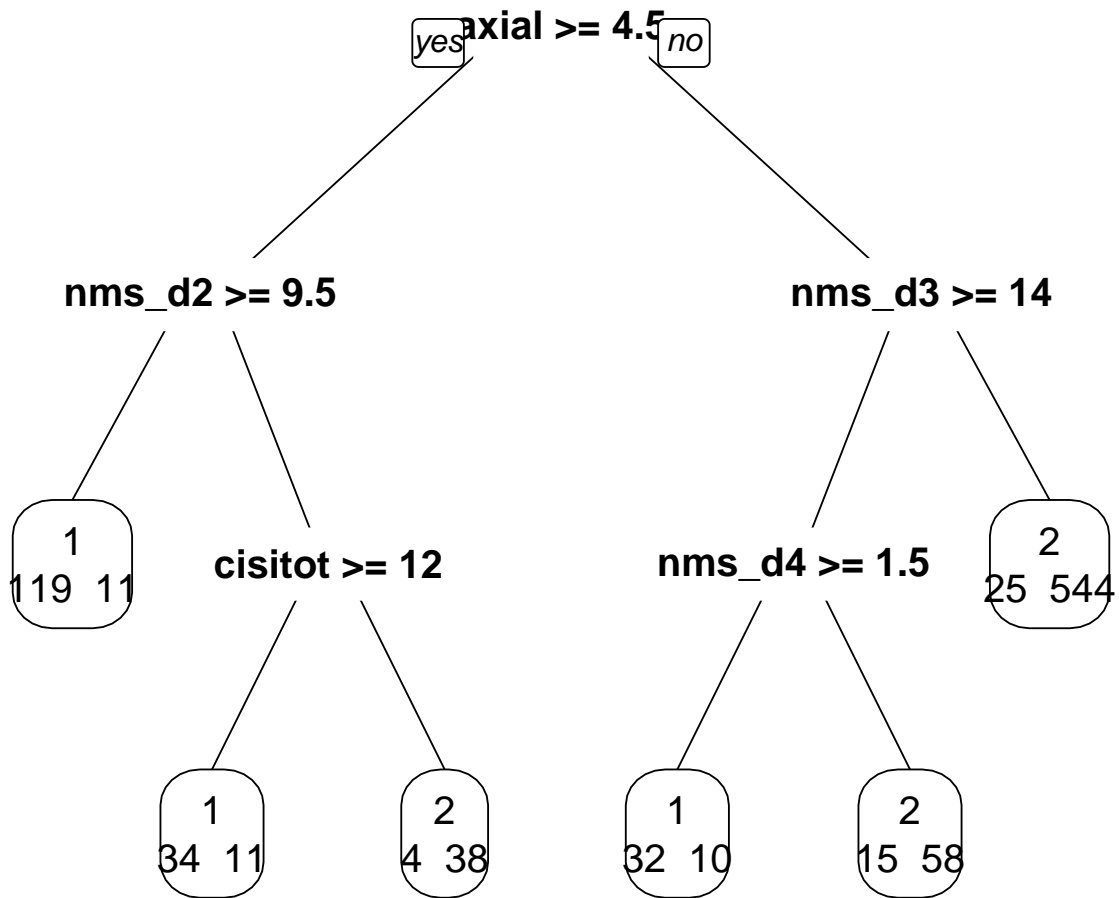


Figure 9: Decision Tree from k -means clustering, 2 clusters

UNSCALED Pruned Tree, 3 clusters

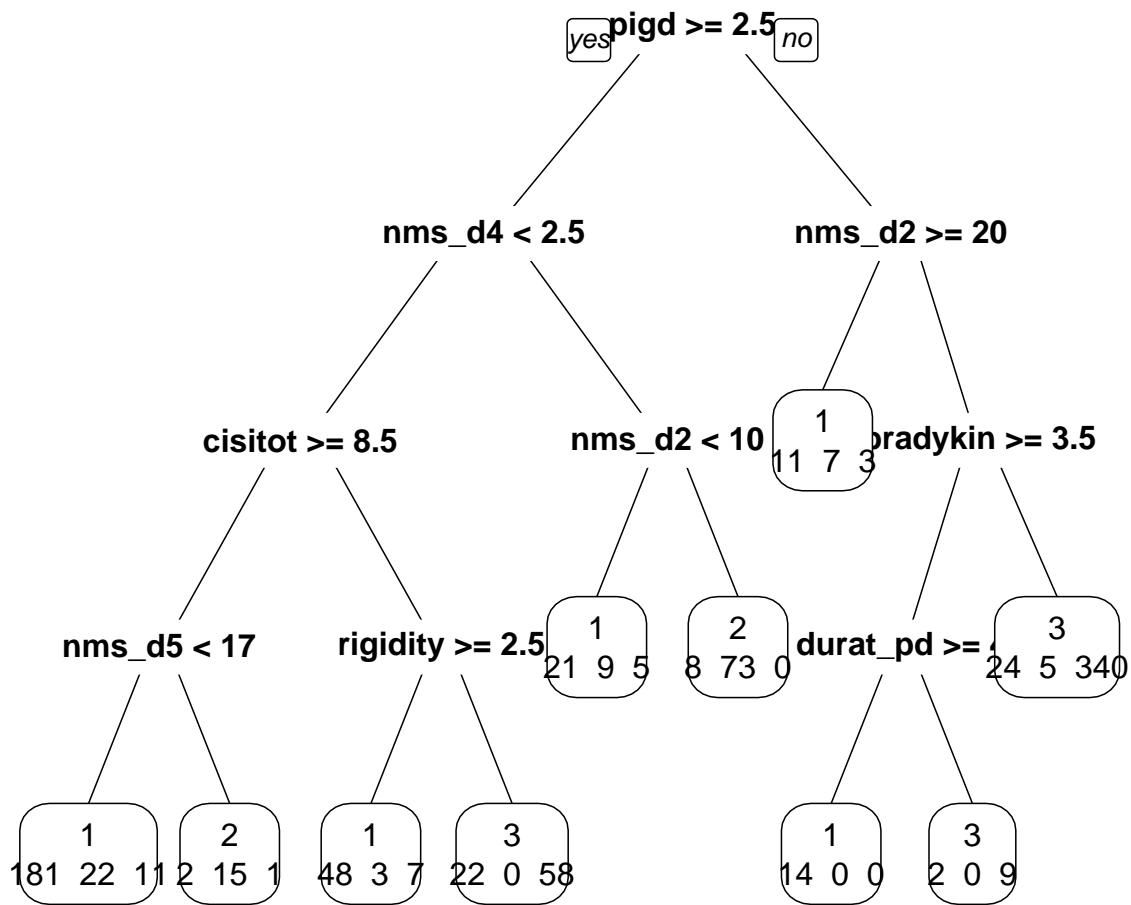


Figure 10: Decision Tree from *k*-means clustering, 3 clusters

UNSCALED Pruned Tree, 4 clusters

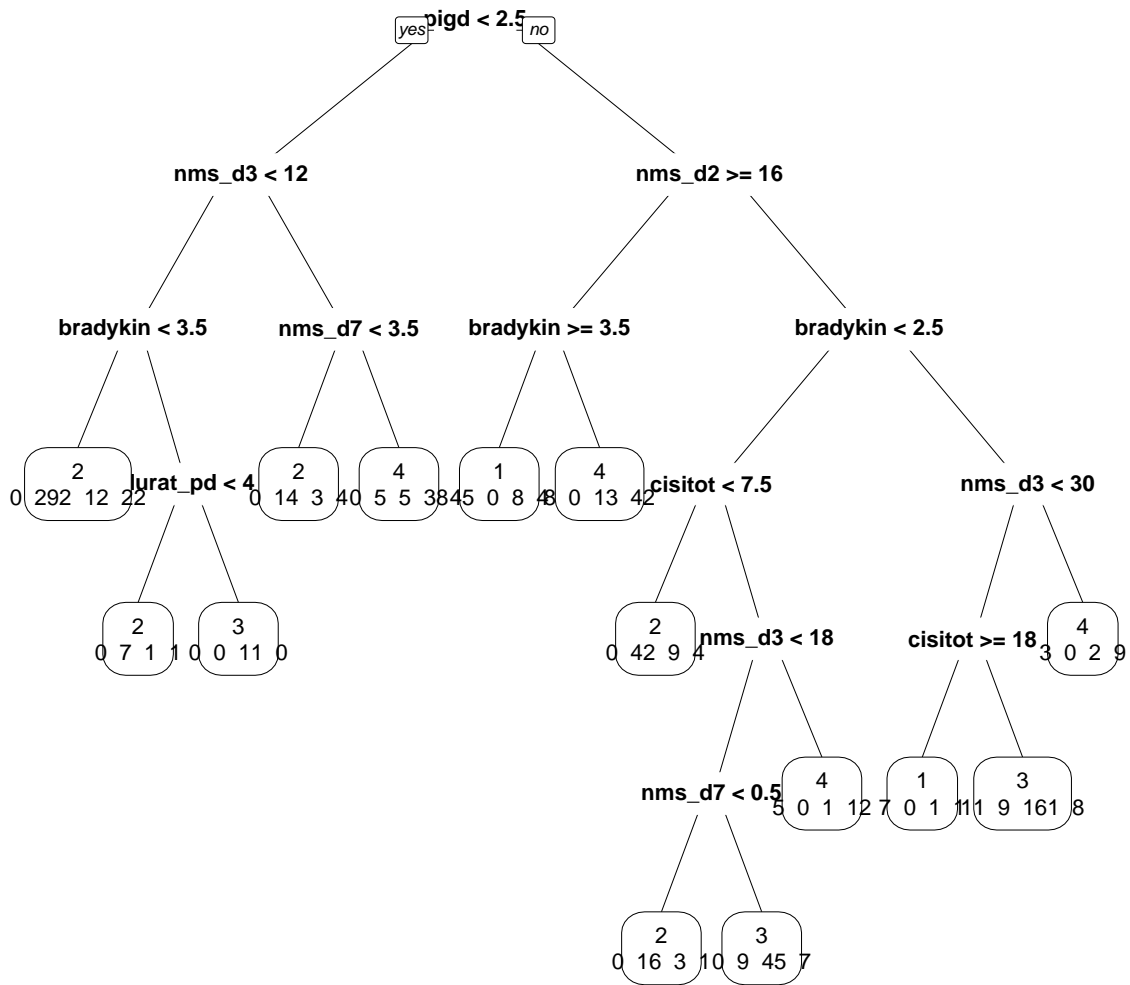


Figure 11: Decision Tree from *k*-means clustering, 4 clusters

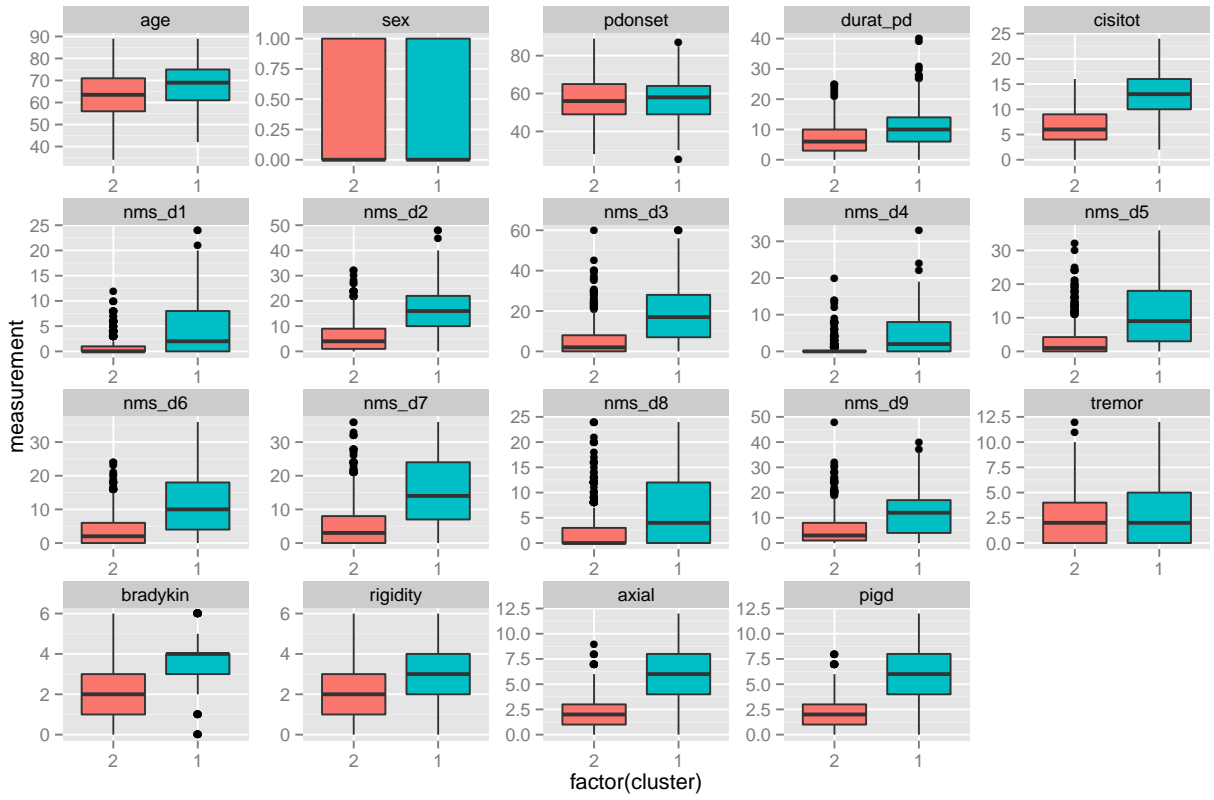


Figure 12: Cluster Summaries, $k = 2$

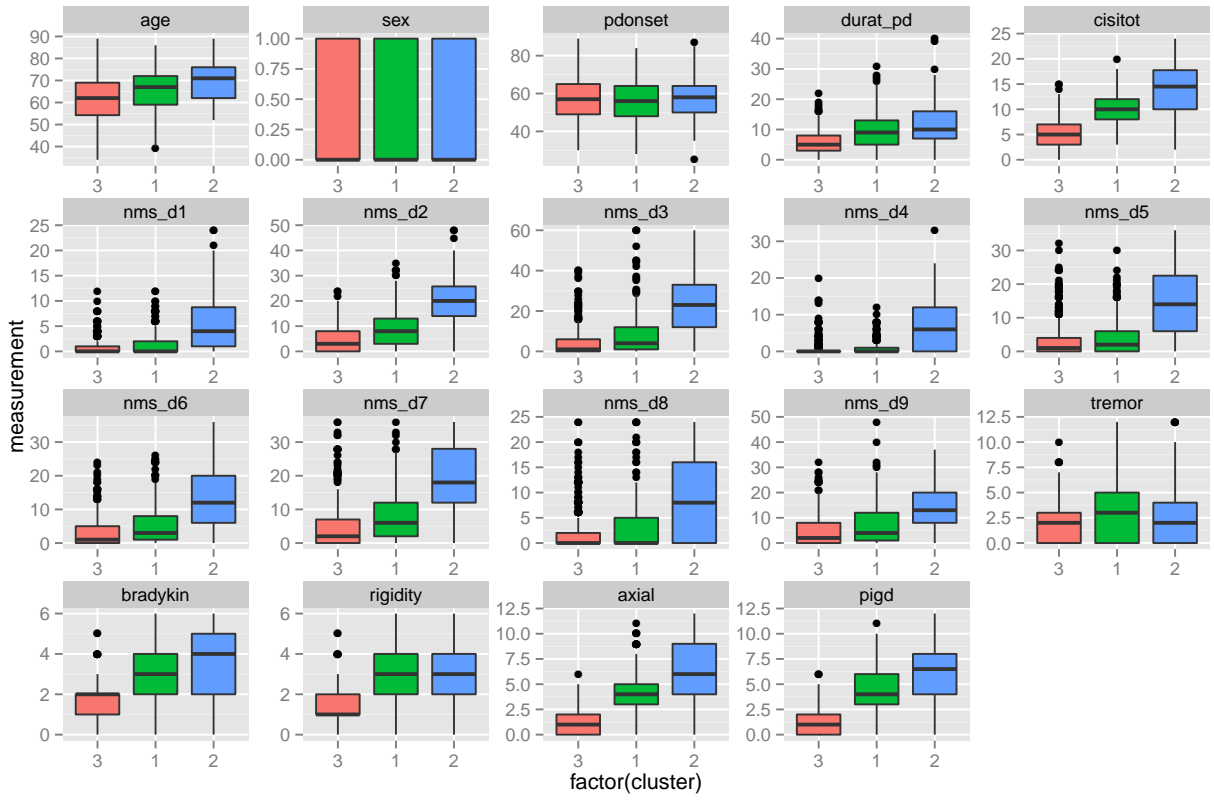


Figure 13: Cluster Summaries, $k = 3$

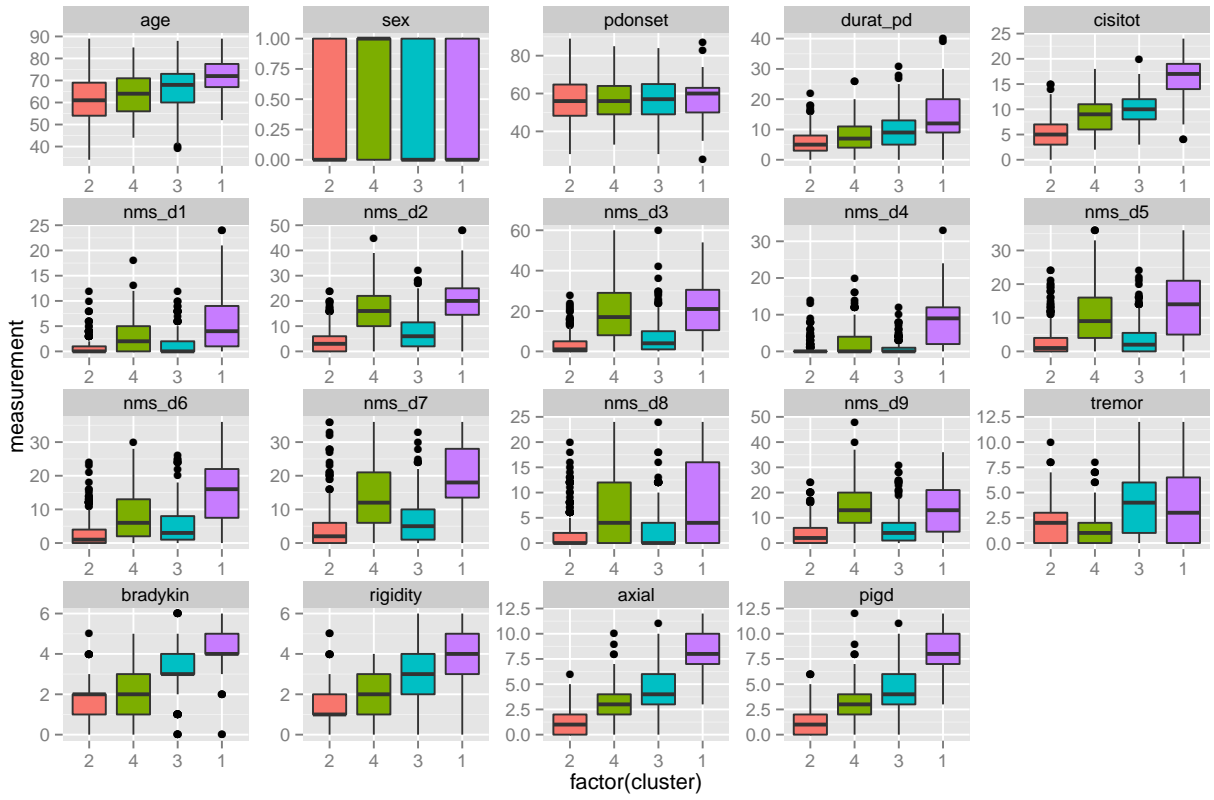


Figure 14: Cluster Summaries, $k = 4$

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.

AP Silhouette Plot $k = 10$

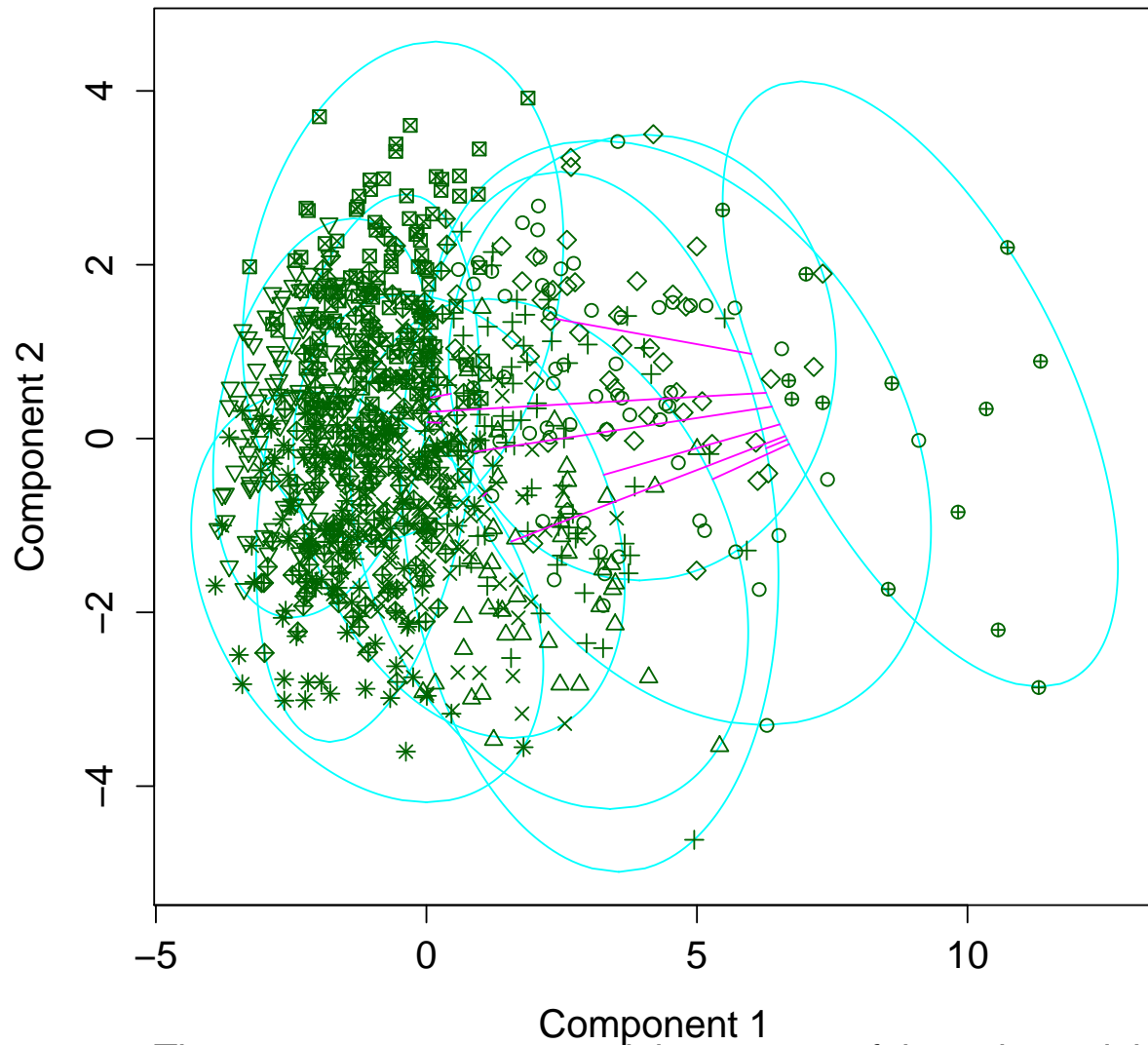


Figure 15: AP silhouette plot, $k = 8$

4 Hierarchical Clustering

4.1 Clustering

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

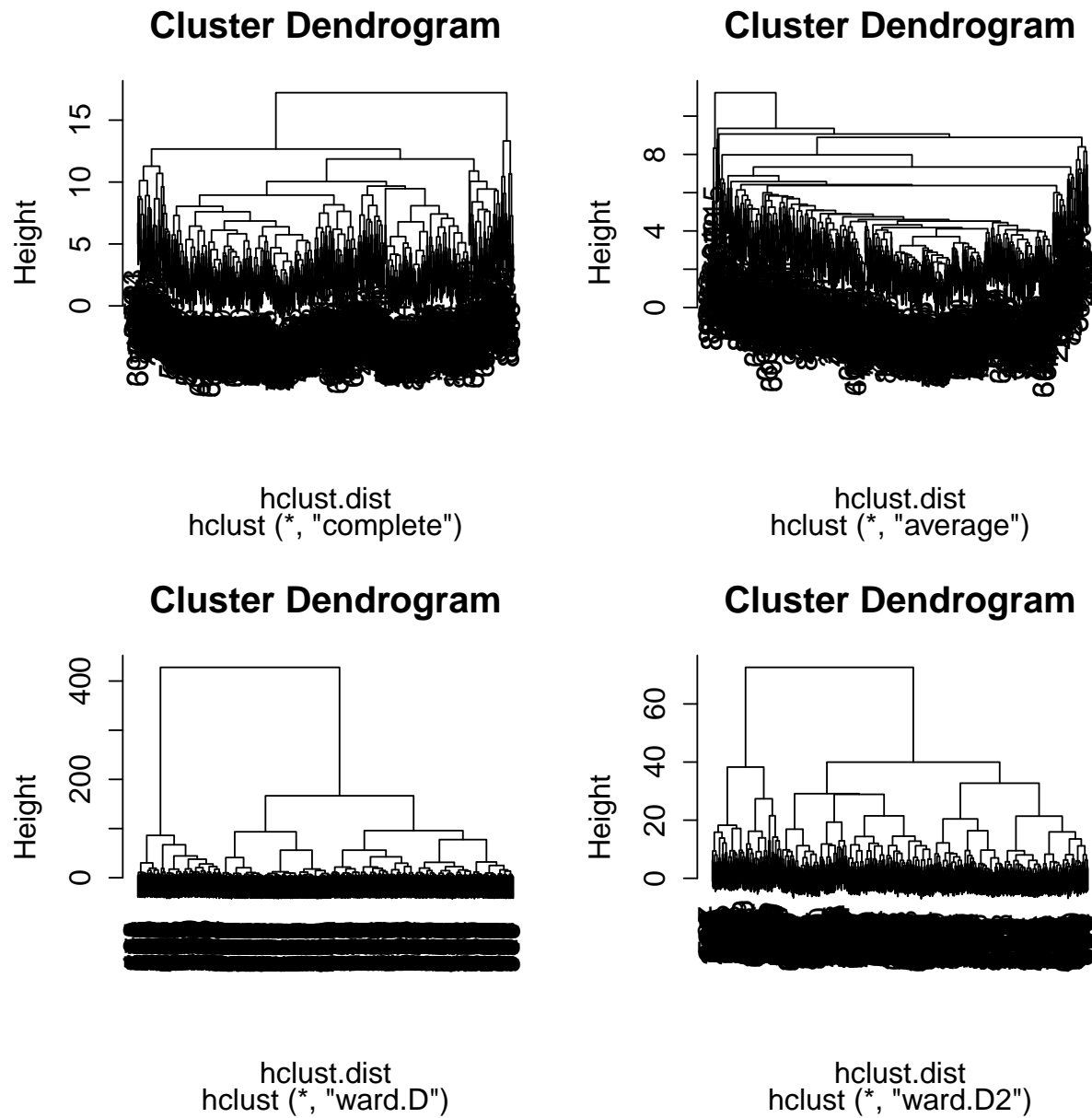


Figure 16: Dendrograms

Method	Condition	n	Figure
Complete	$k = 4$	4 (790/81/18/12)	17
Complete	<code>dynamicTreeCut</code> 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 6: Clusters from Tree Cutting

4.2 Cutting Trees

4.3 Interpretation

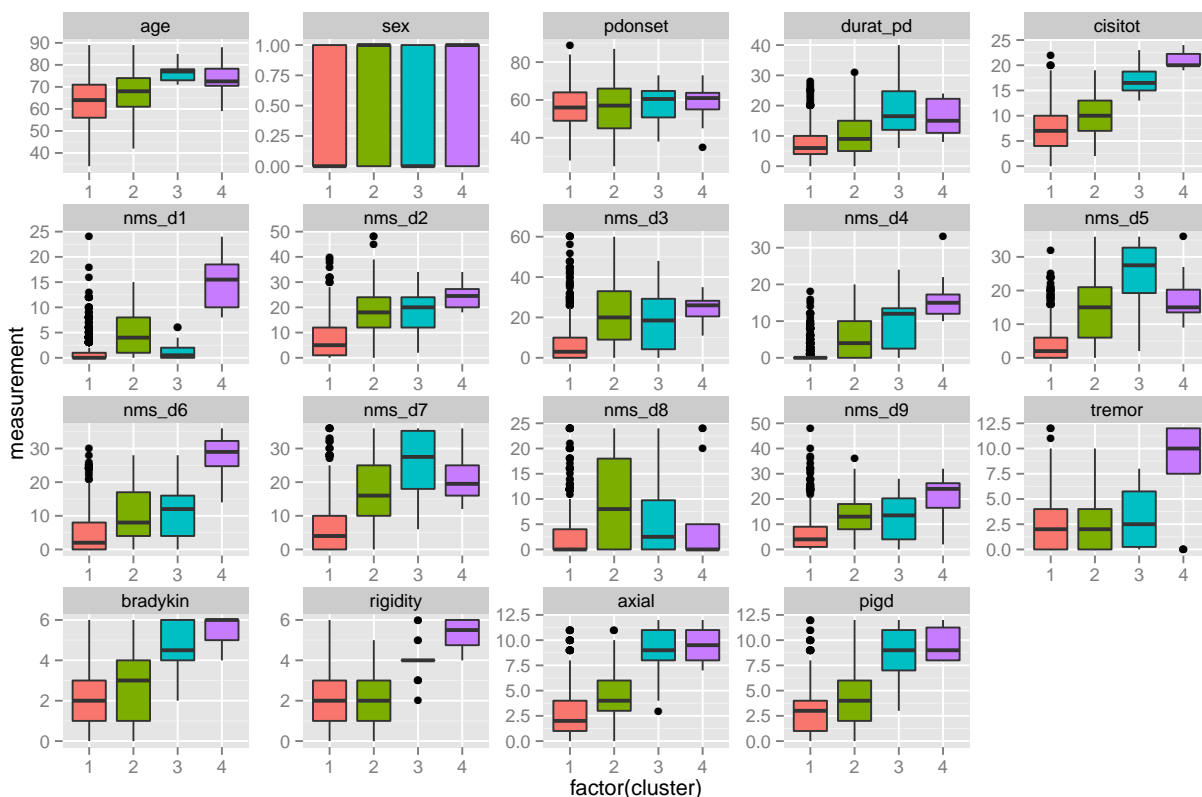


Figure 17: Using maximum (complete linkage) dissimilarity, cutting tree for $k = 4$

4.4 Interpretation

Cluster sizes are available in Table 7

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

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

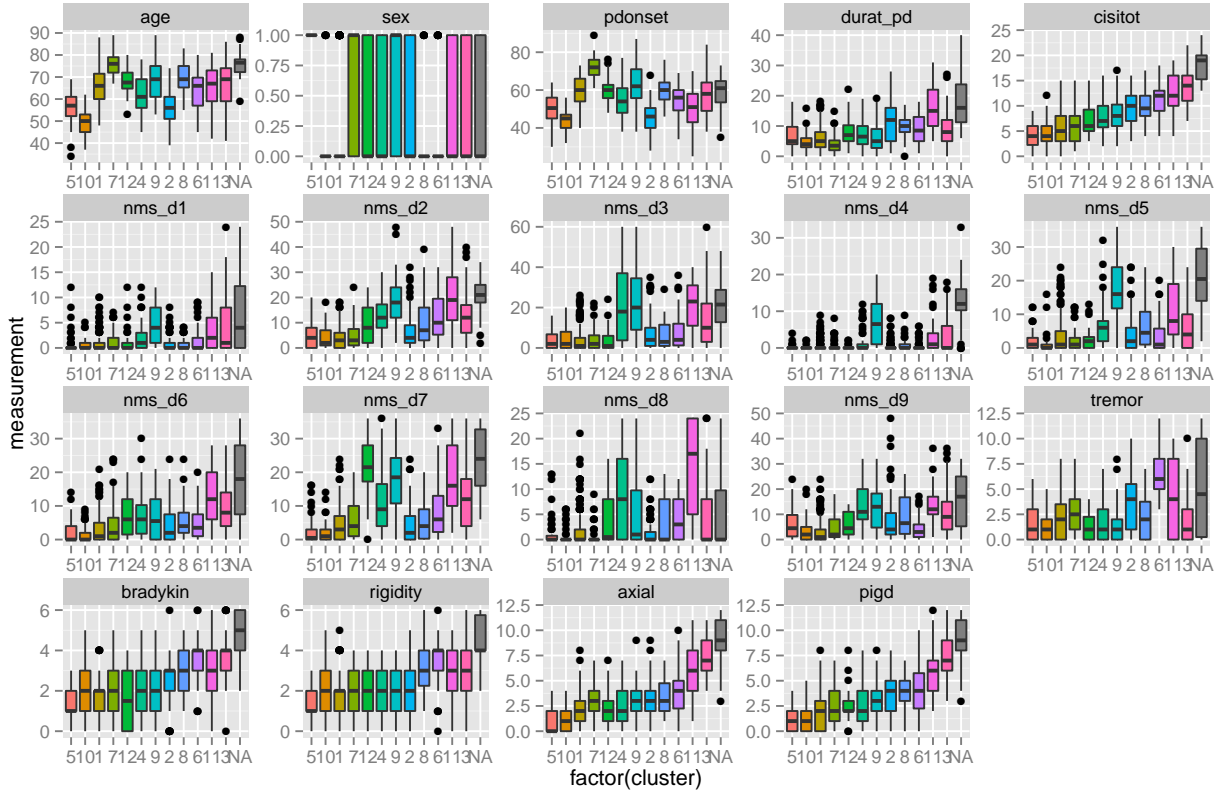


Figure 18: Using maximum (complete linkage) dissimilarity, cutting tree dynamically

5 Biclustering

Used BCBimax clustering algorithm. Clusters seem quite sparse.

6 Subspace clustering

7 Bayesian Networks

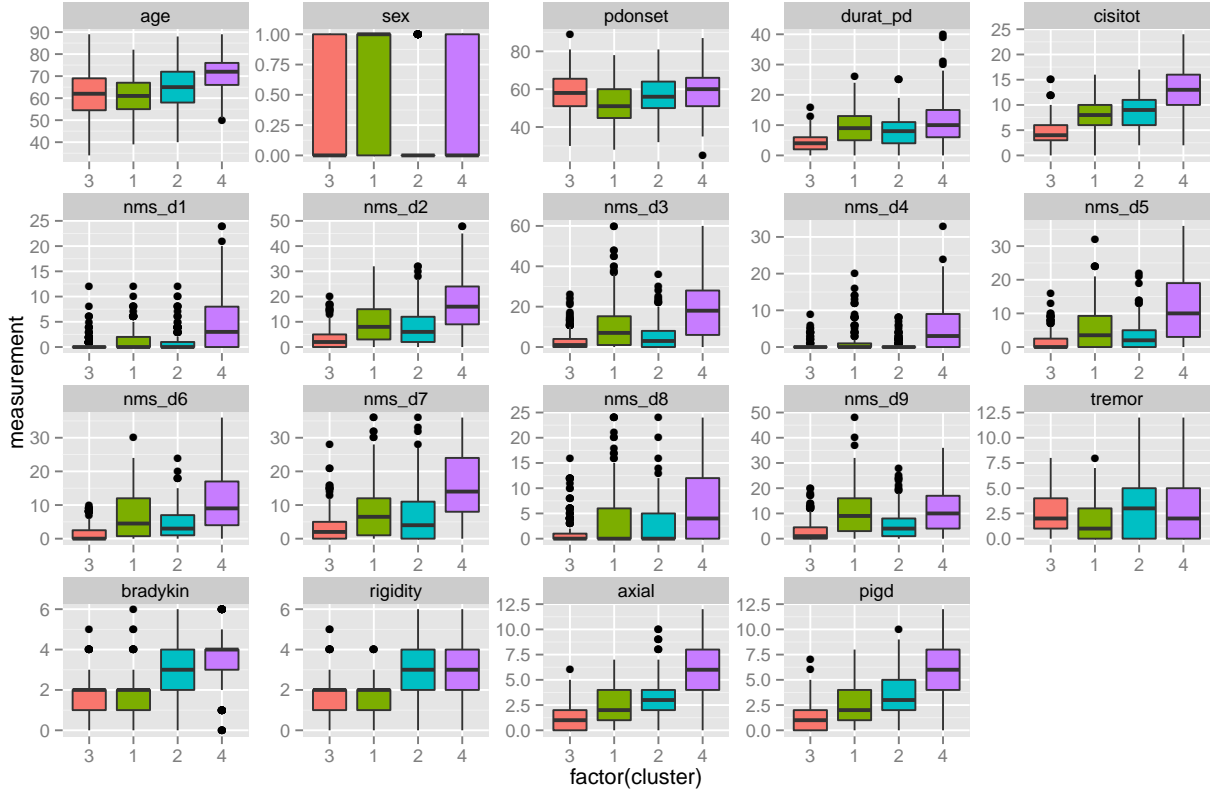


Figure 19: Using Ward (1963) dissimilarity, cutting tree for $k = 4$

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

Table 7: AP Cluster Sizes

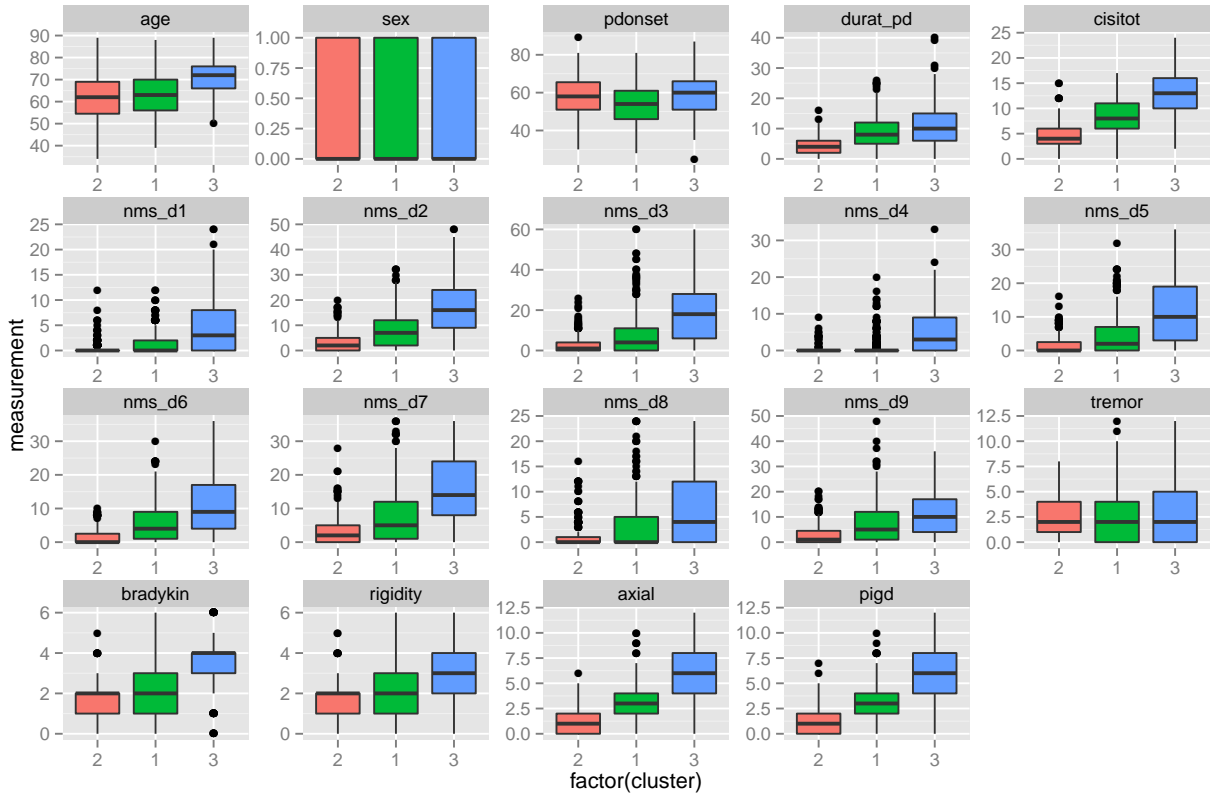


Figure 20: Using Ward (1963) dissimilarity, cutting tree at $h = 100$

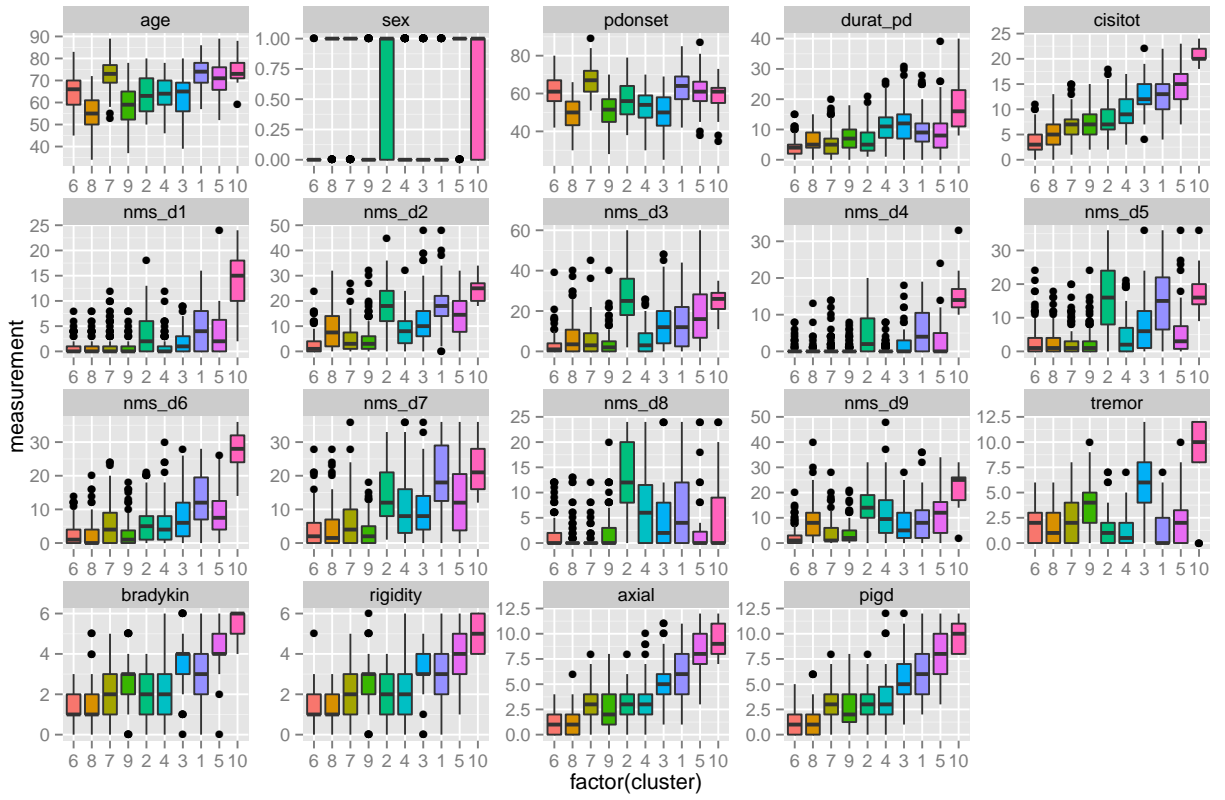


Figure 21: AP Boxplot Summaries

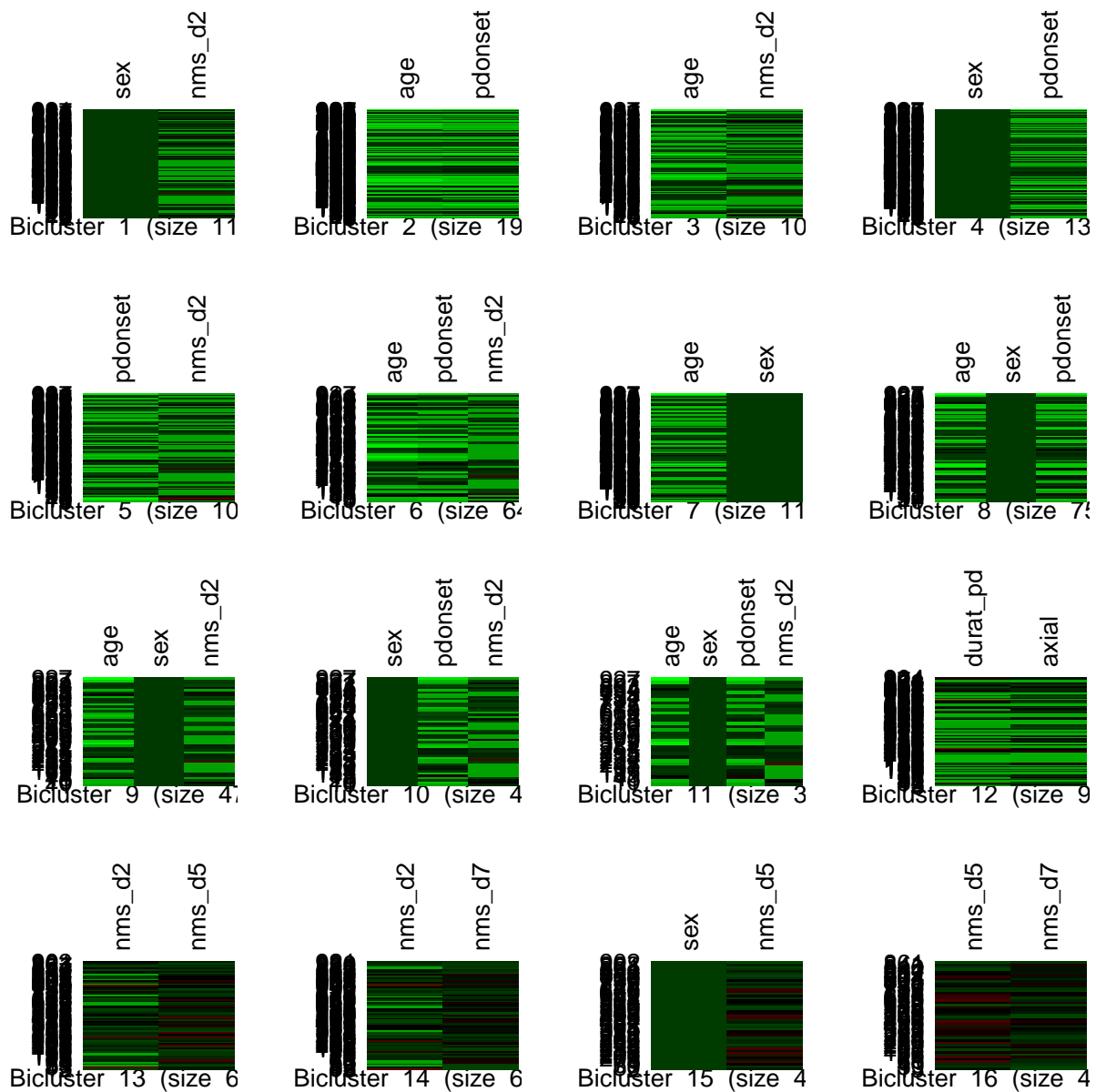


Figure 22: Biclustering heatmaps $N = 16$

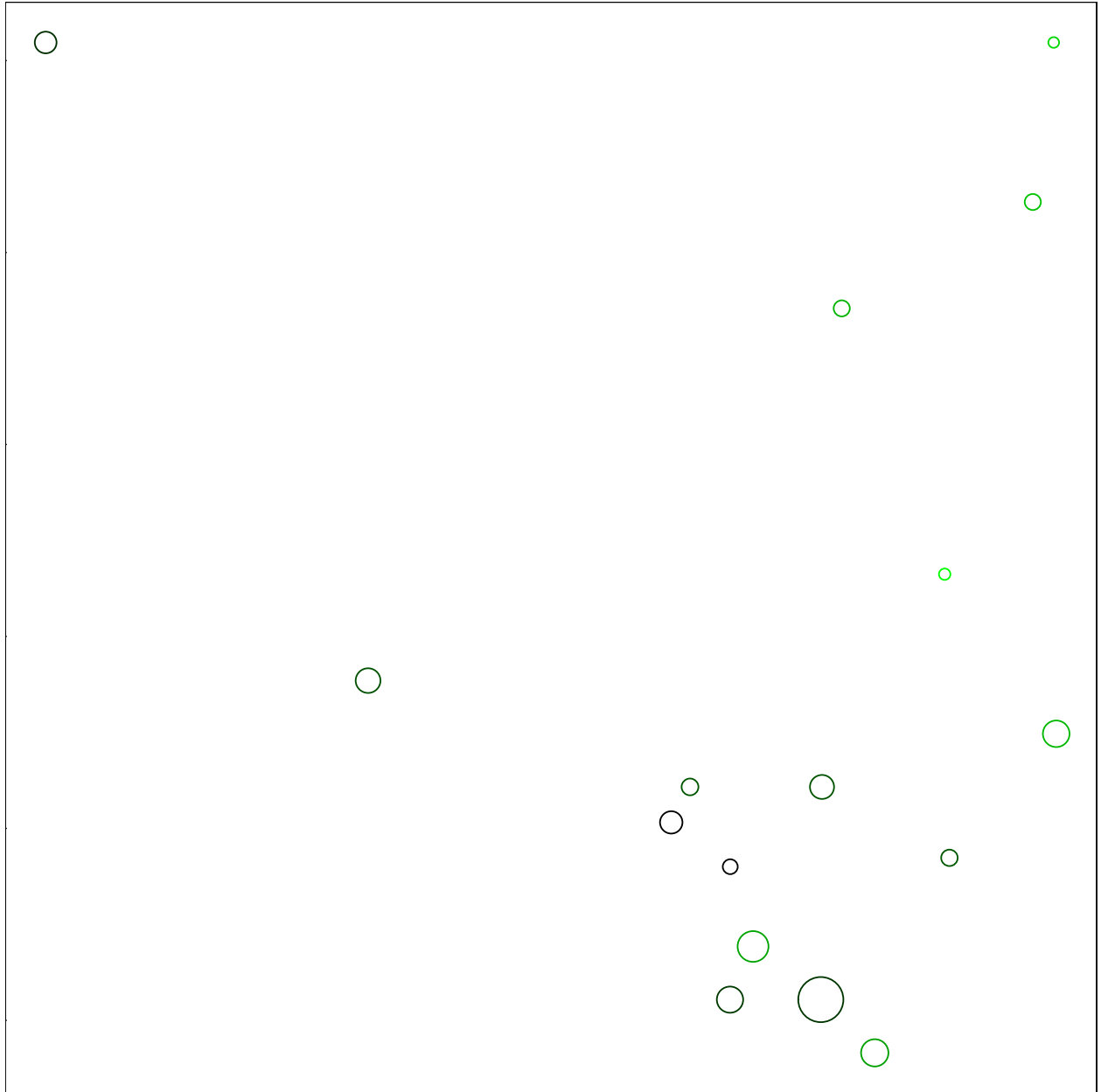


Figure 23: Bubbleplot $N = 16$