

# 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 <sup>1</sup>
rigidity	float	%9.0g	rigidity
axial	float	%9.0g	axial <sup>2</sup>
pigd	float	%9.0g	postural instability and gait difficulty

Table 1: Selected Features and Details

---

<sup>1</sup>Impaired ability to adjust the body's position.

<sup>2</sup>Issues affecting the middle of the body.

Name	$\mu$	$\sigma$	min-max
nms_d1	1.76	3.32	0-24
nms_d2	8.71	8.76	0-48
nms_d3	8.70	11.83	0-60
nms_d4	1.65	3.94	0-33
nms_d5	5.22	7.44	0-36
nms_d6	5.67	6.92	0-36
nms_d7	8.02	9.09	0-36
nms_d8	3.57	5.97	0-24
nms_d9	6.99	7.74	0-48
tremor	2.59	2.63	0-12
bradykin	2.49	1.39	0-6
rigidity	2.34	1.36	0-6
axial	3.28	2.75	0-12
pigd	3.36	2.77	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.

Figure 1 shows scree test elbow occurs around 2 or 3. Also, eigenvalues 1 and 2  $> 1$ , while 3 is around .9

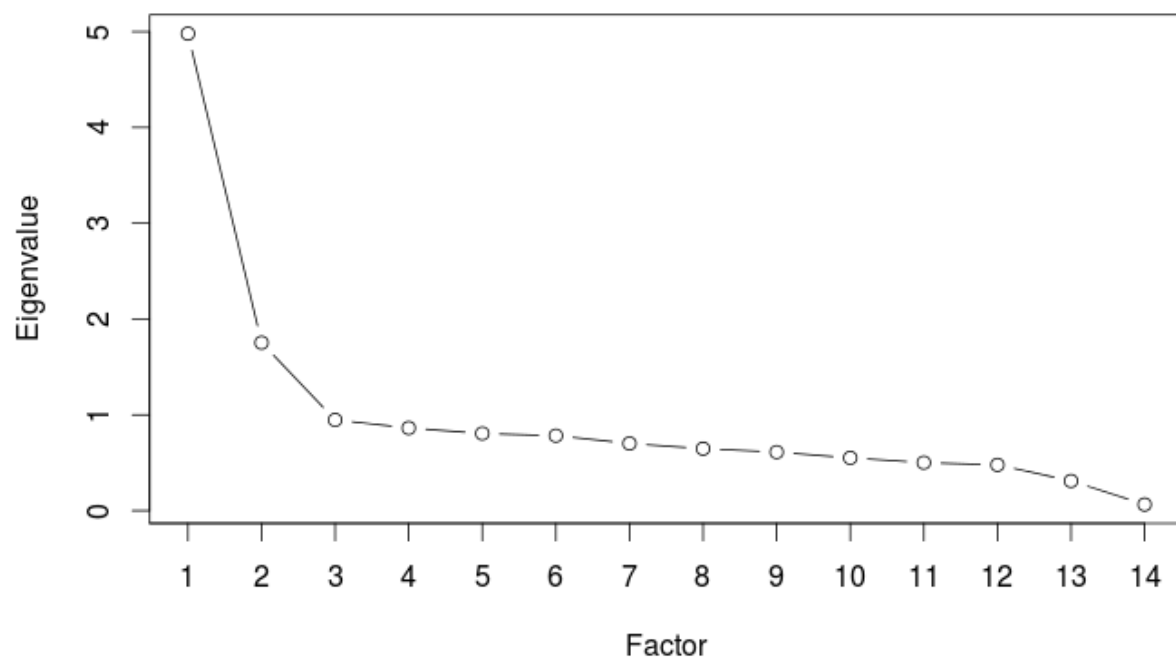


Figure 1: Scree test: eigenvalues by factor

## 2 $k$ -means

### 2.1 Identifying optimal number of clusters

#### 2.1.1 WSS Error Scree Test

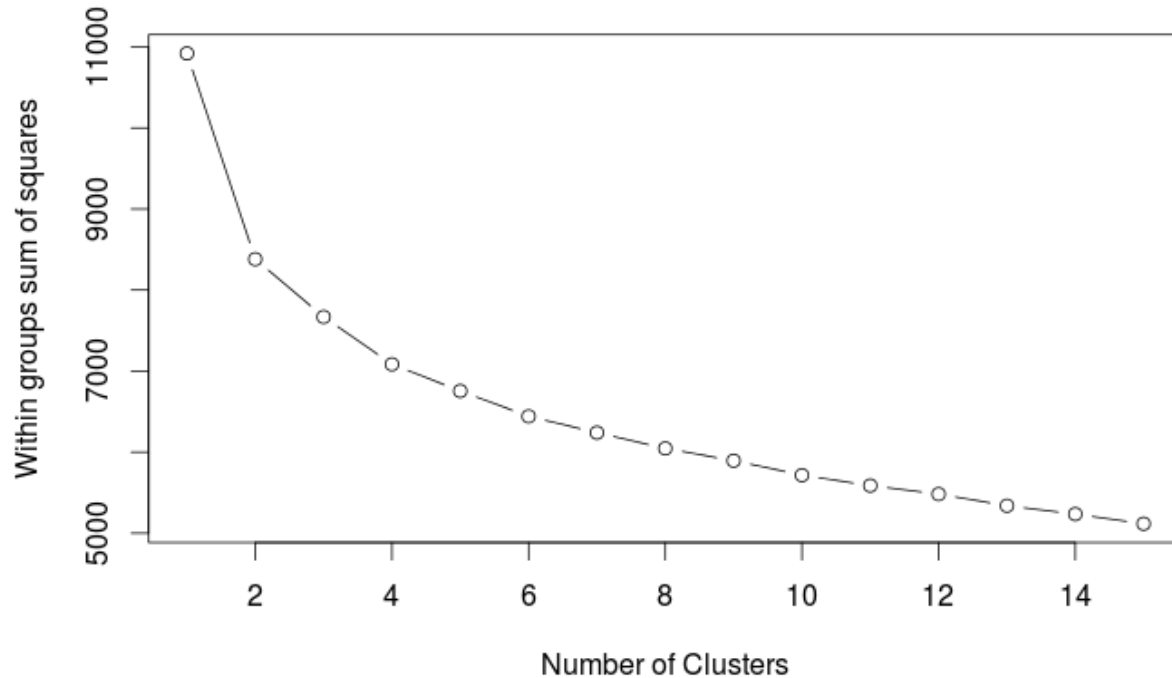


Figure 2: Scree test: WSS error by cluster size

Figure 2 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 3.

#### 2.1.3 Average Silhouette Width

Figure 4 shows average silhouette width as being consistently under 0.25 for all clusters, implying the data is not well structured.

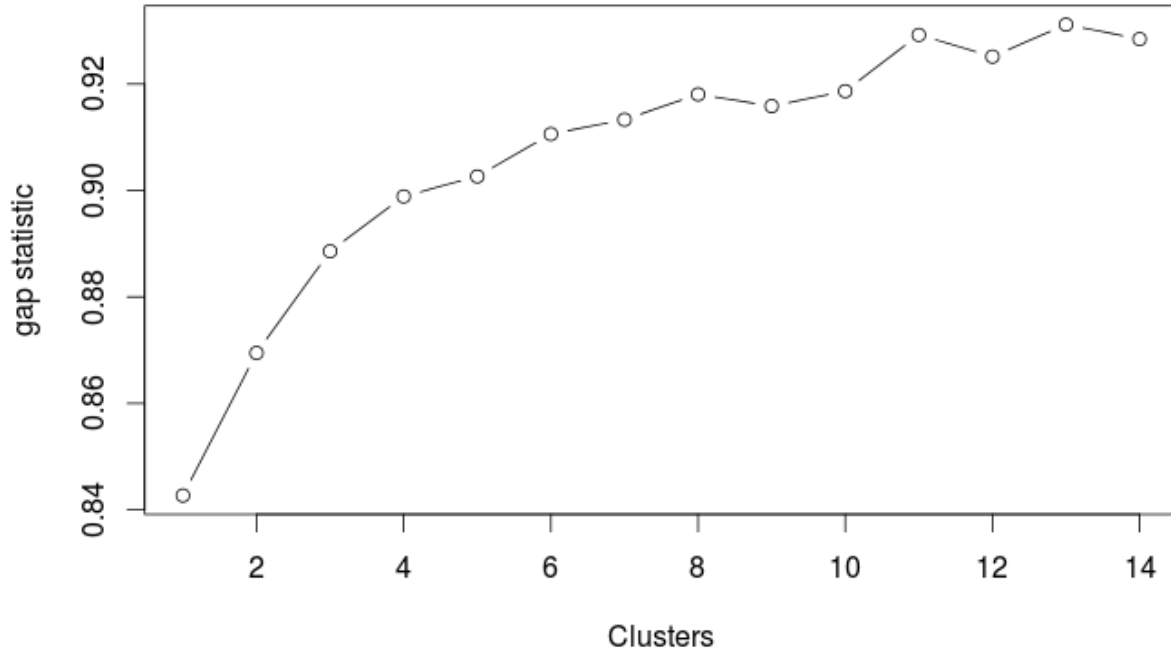


Figure 3: Gap statistic by cluster size

$k$	$n$	Within SS	sum(Within SS)
2	201/580	4248.585/4132.434	8381.019
3	420/231/130	2618.368/1973.82/3076.542	7668.73
4	61/372/145/203	1481.25/1845.389/2147.988/1609.555	7084.183

Table 3: Cluster statistics

## 2.2 Cluster statistics

## 2.3 Silhouette plots

Available in Figures 5, 6, and 7. Note: constructed with standardized  $z$ -score data.

## 2.4 Decision trees based on clusters

## 2.5 Interpretation of Clusters

### 2.5.1 Cluster summaries

Available in Figures 11, 12, and 13. Error bar is standard error.

---

<sup>3</sup>Complexity Parameter

<sup>4</sup>10-fold cross validation

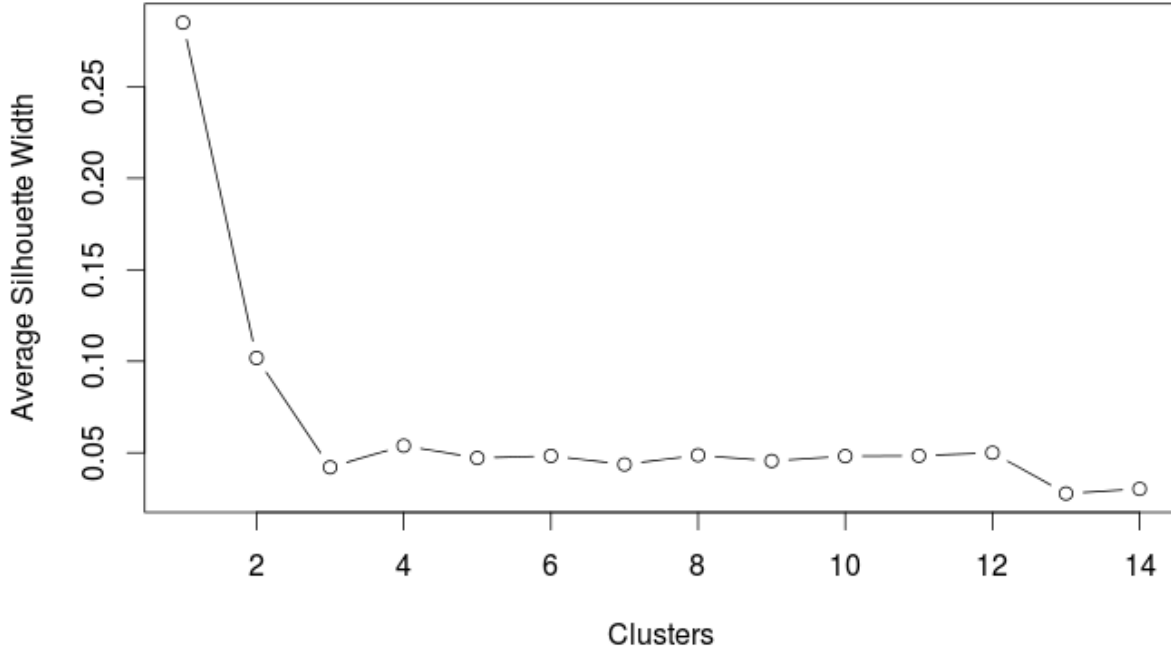


Figure 4: Average silhouette width by cluster size

$k$	CP <sup>3</sup>	CV Xerror <sup>4</sup>	Root Feature	Root Error	Figure
2	0.0348	0.134	axial $\geq 0.44$	0.257	Figure 8
3	0.0100	0.194	bradykin $< 0.0041$	0.462	Figure 9
4	0.0100	0.248	bradykin $< 0.0041$	0.523	Figure 10

Table 4:  $k$ -kmeans decision trees statistics

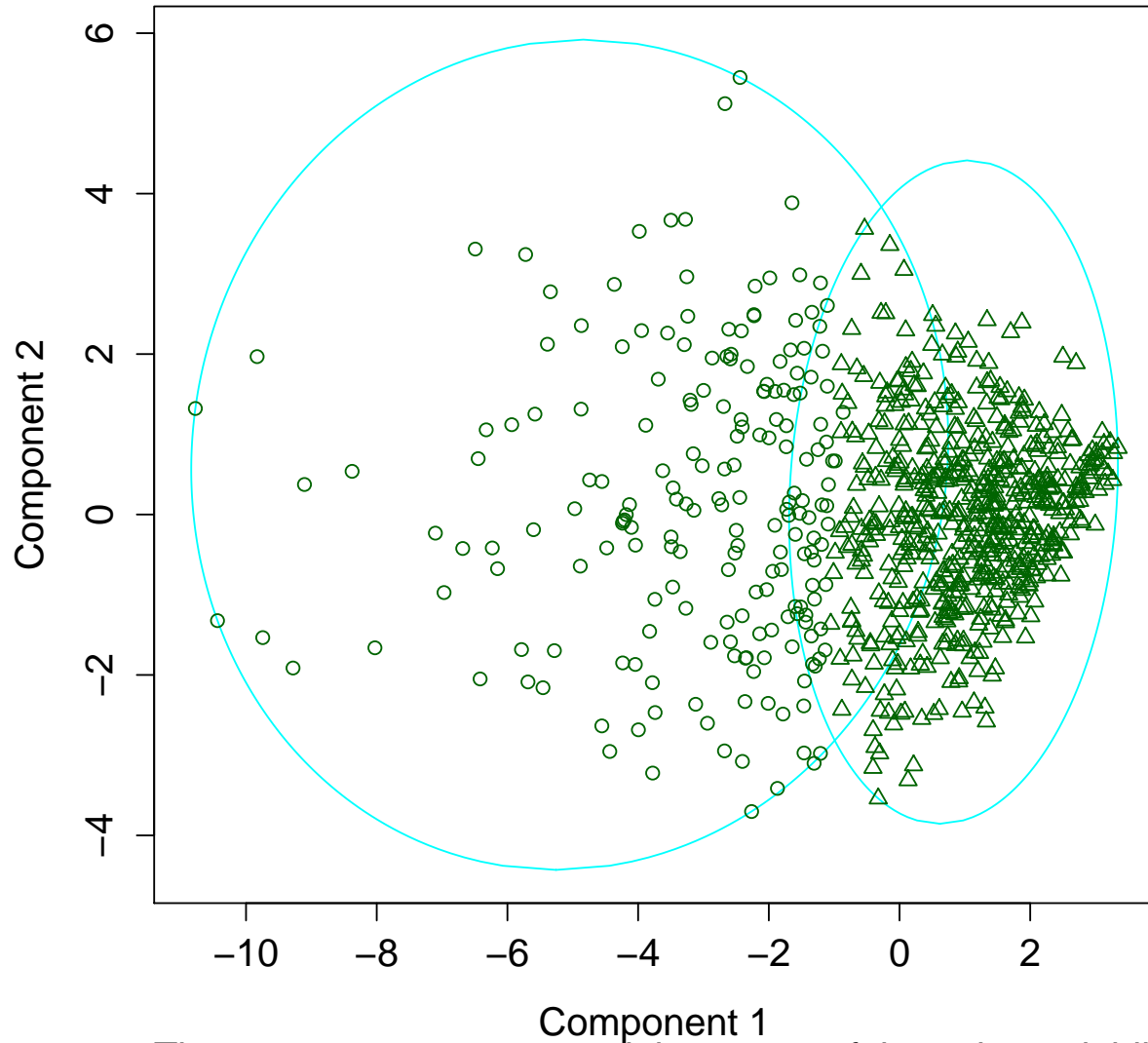
### 2.5.2 Interpretation

$k = 2$  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.

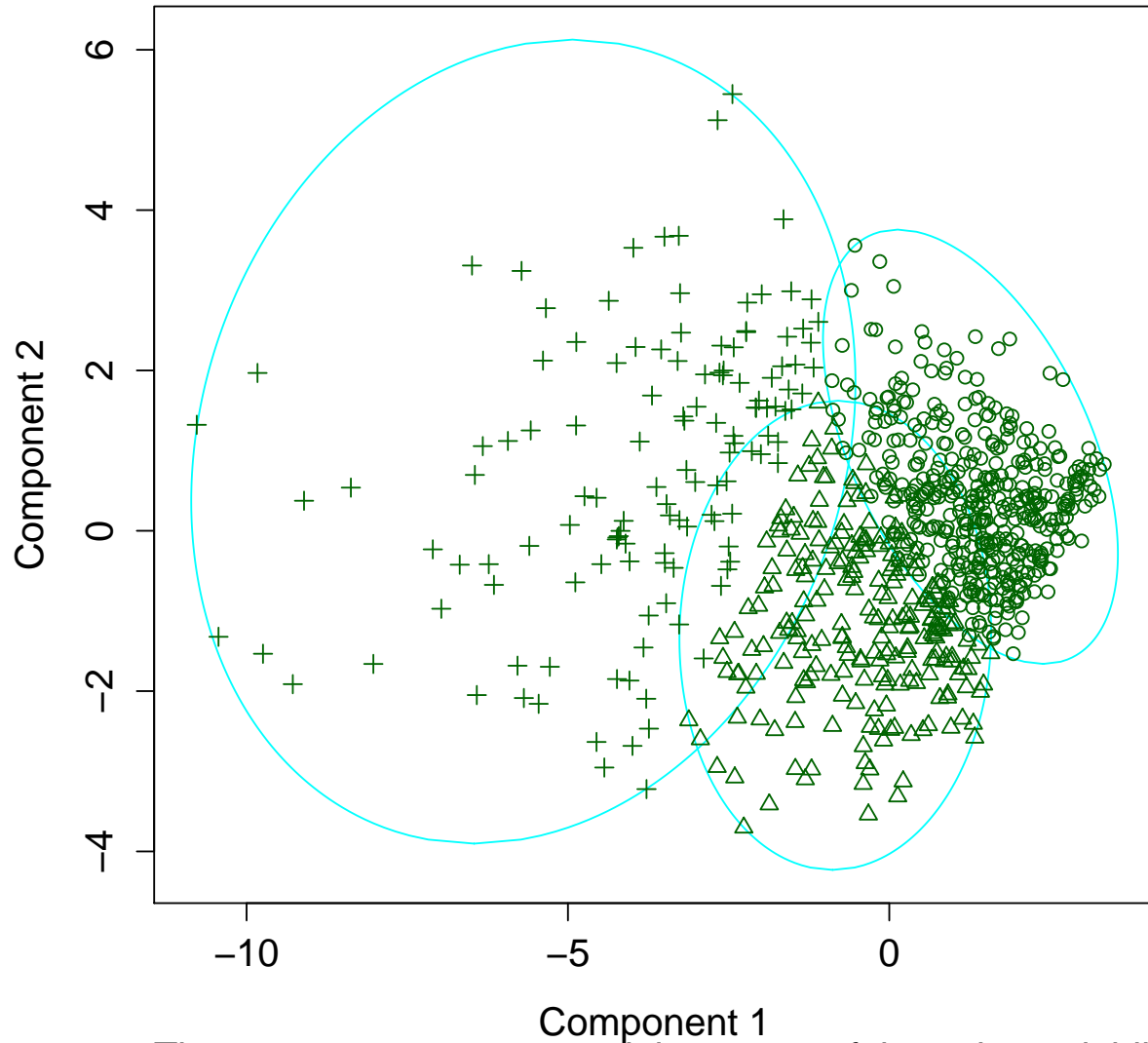
### Silhouette plot $k = 4$



Component 1  
These two components explain 48.1 % of the point variability.

Figure 5:  $k$ -means cluster silhouette plot,  $k = 2$

### Silhouette plot $k = 4$

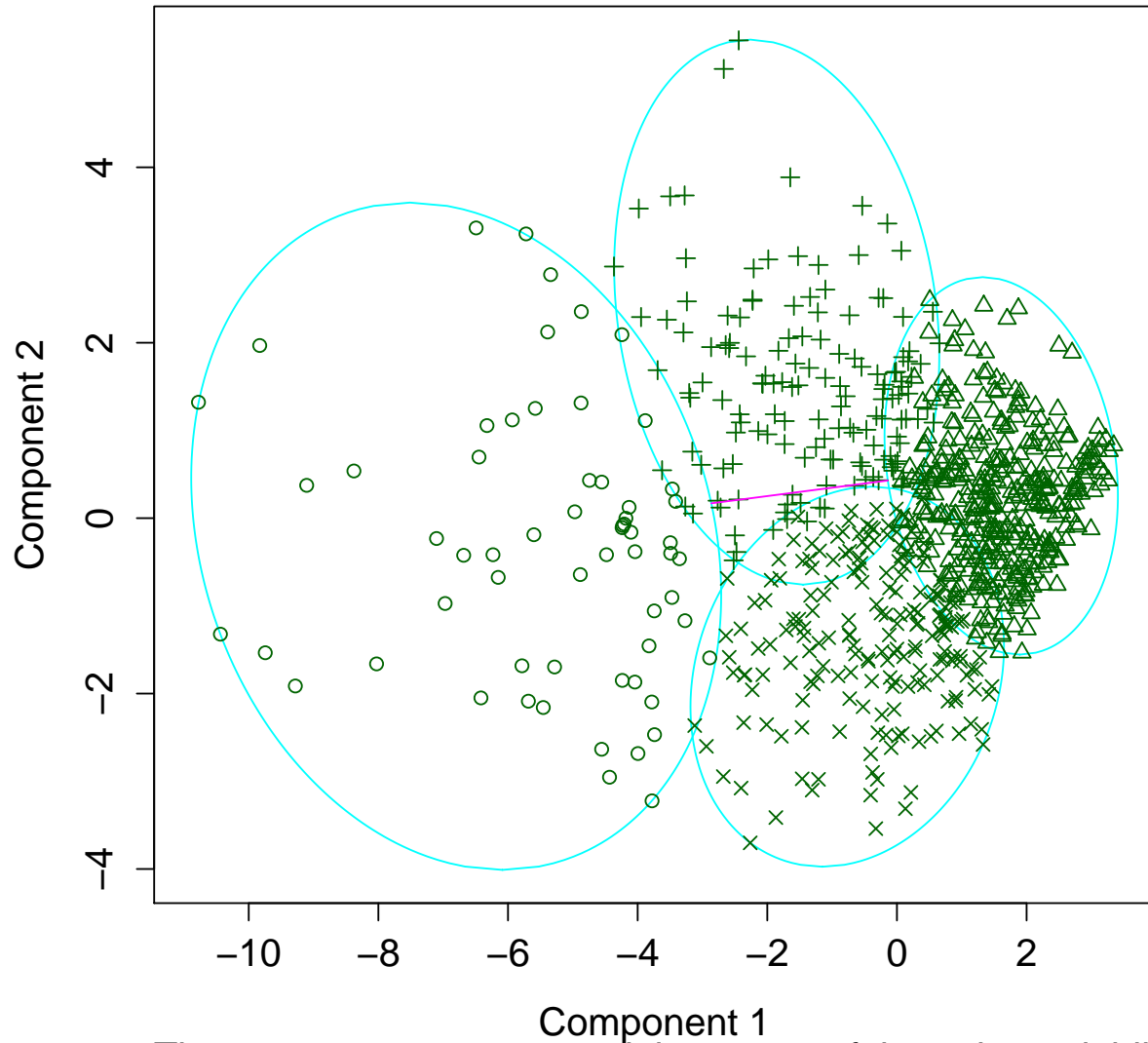


Component 1  
These two components explain 48.1 % of the point variability.

Figure 6:  $k$ -means cluster silhouette plot,  $k = 3$



## Silhouette plot $k = 4$



Component 1  
These two components explain 48.1 % of the point variability.

Figure 7:  $k$ -means cluster silhouette plot,  $k = 4$

## UNSCALED Pruned Tree, 2 clusters

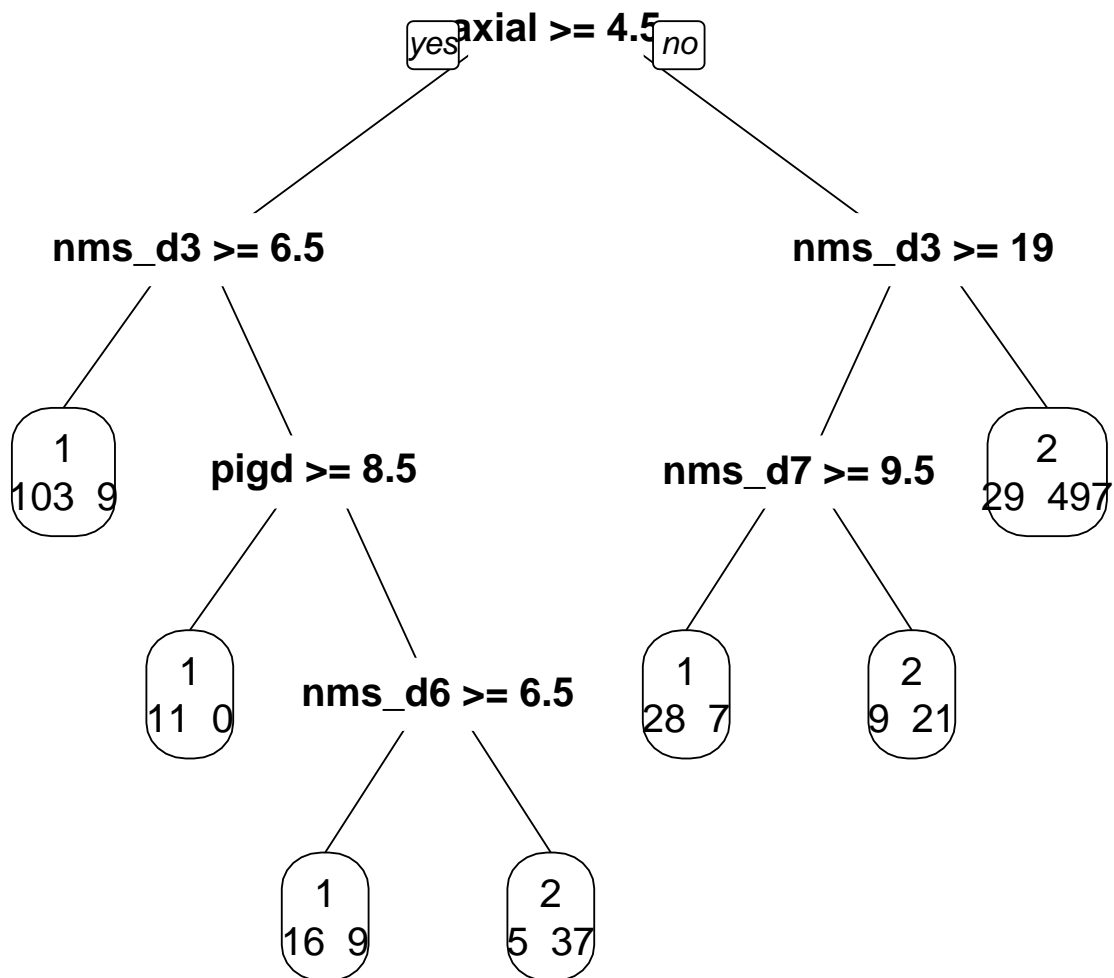


Figure 8: Decision Tree from  $k$ -means clustering, 2 clusters

## UNSCALED Pruned Tree, 3 clusters

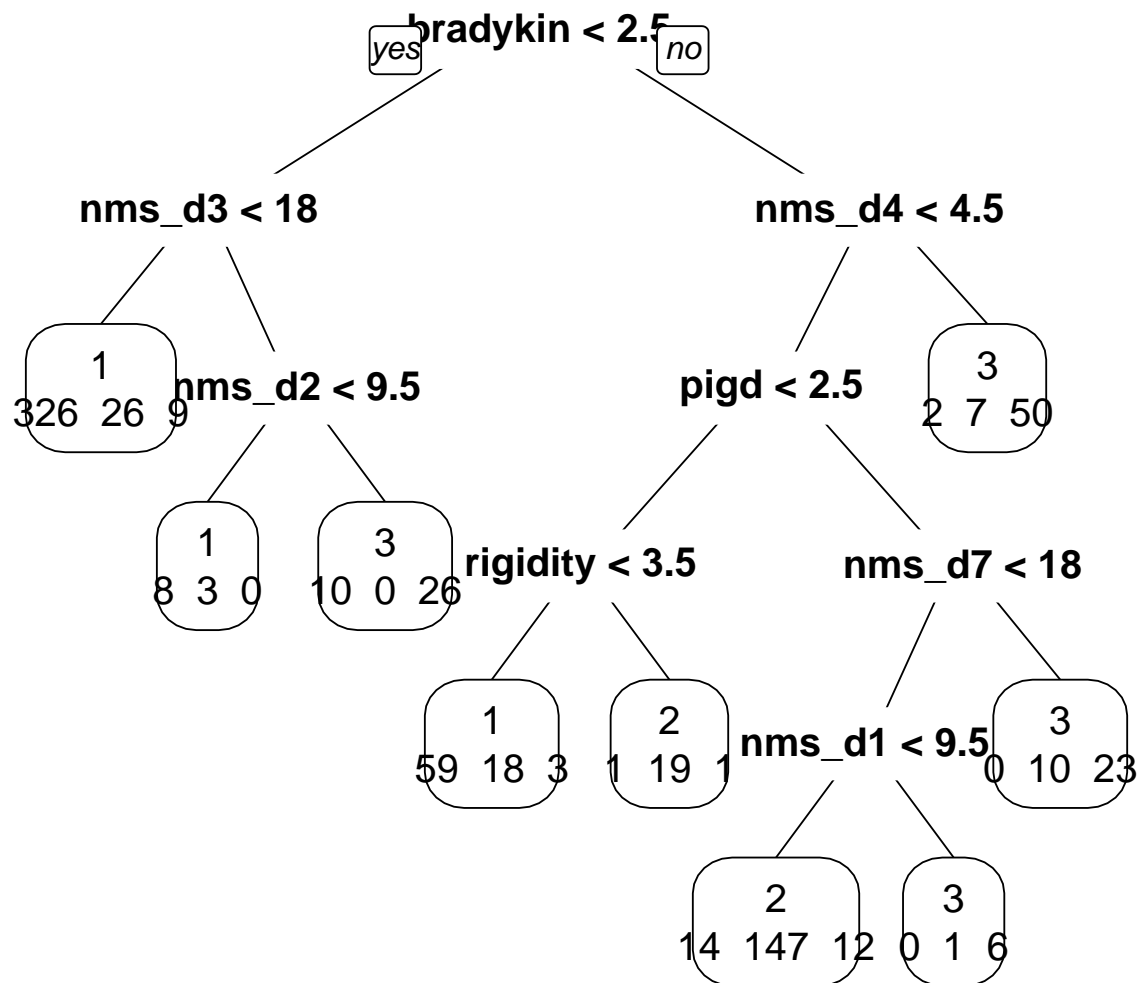


Figure 9: Decision Tree from  $k$ -means clustering, 3 clusters

## UNSCALED Pruned Tree, 4 clusters

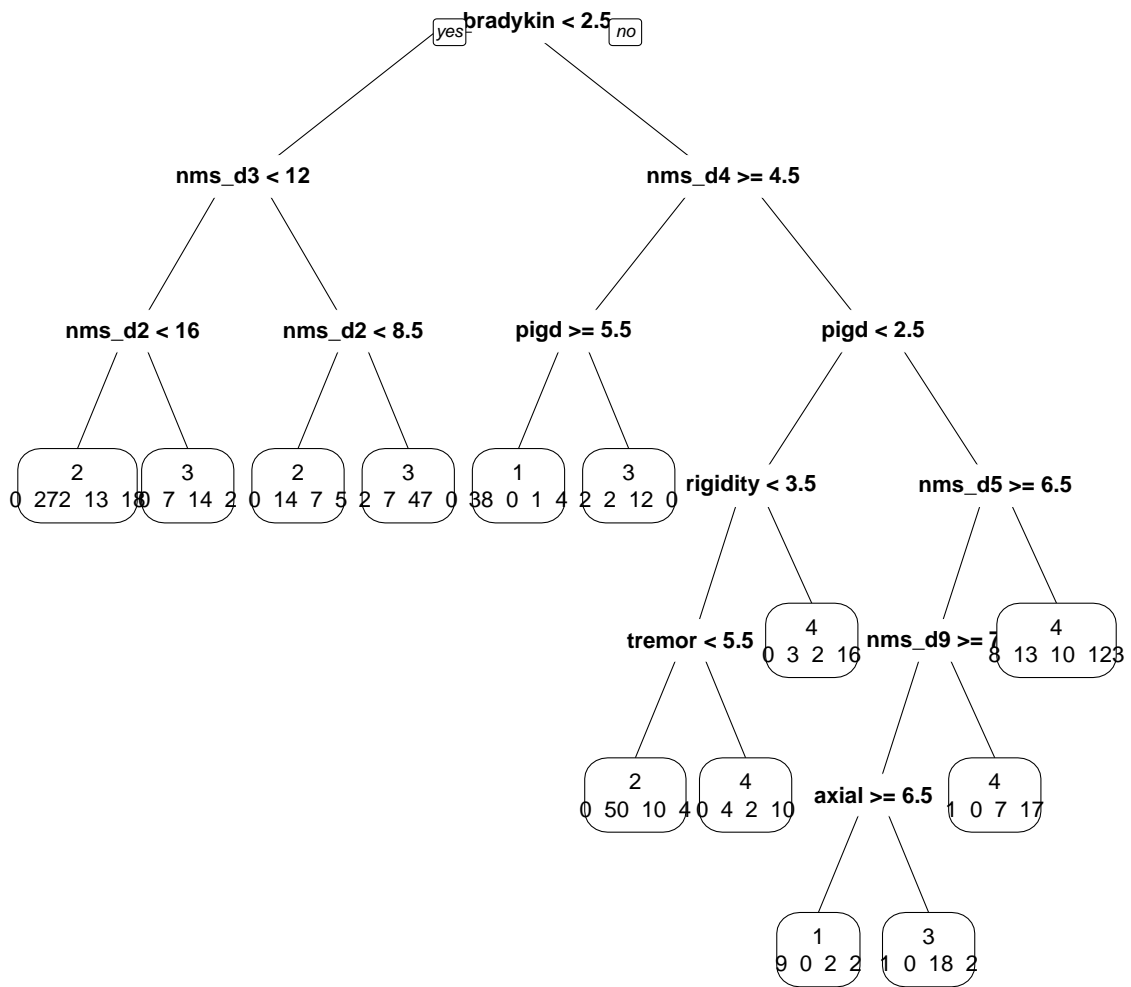


Figure 10: Decision Tree from  $k$ -means clustering, 4 clusters

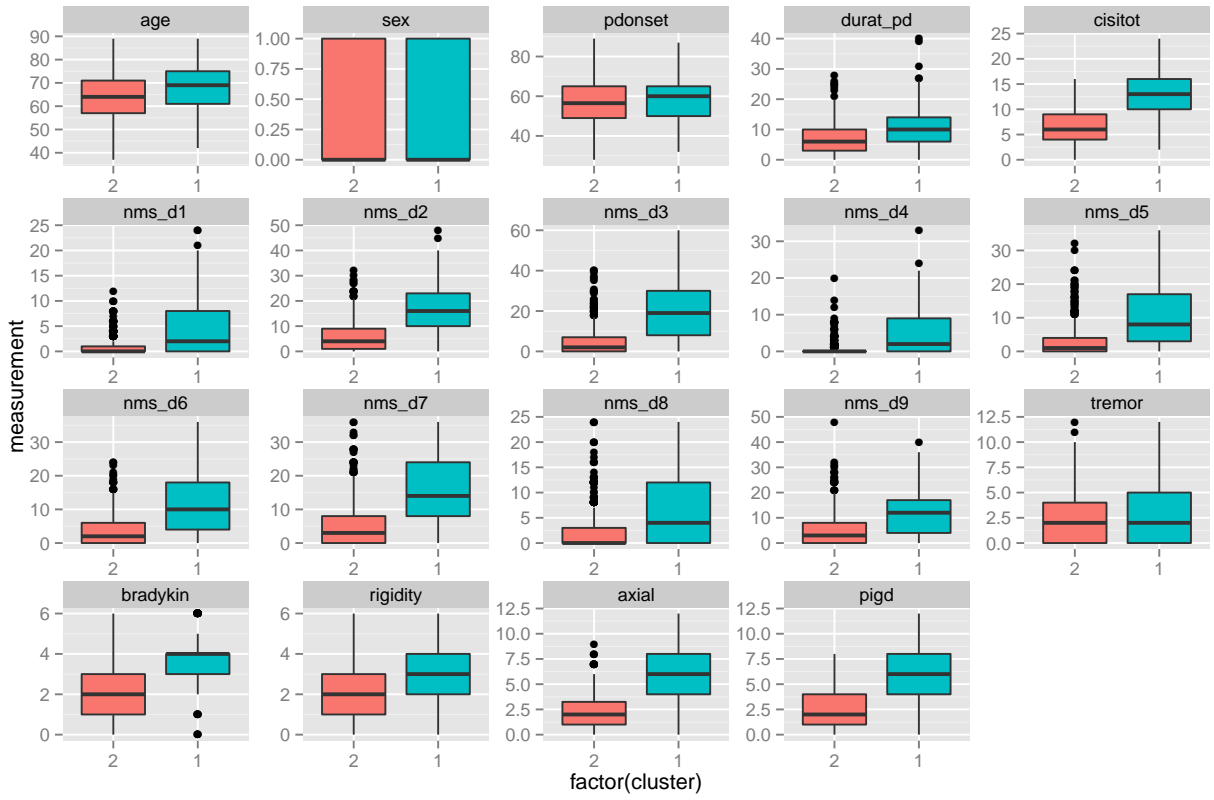


Figure 11: Cluster Summaries,  $k = 2$

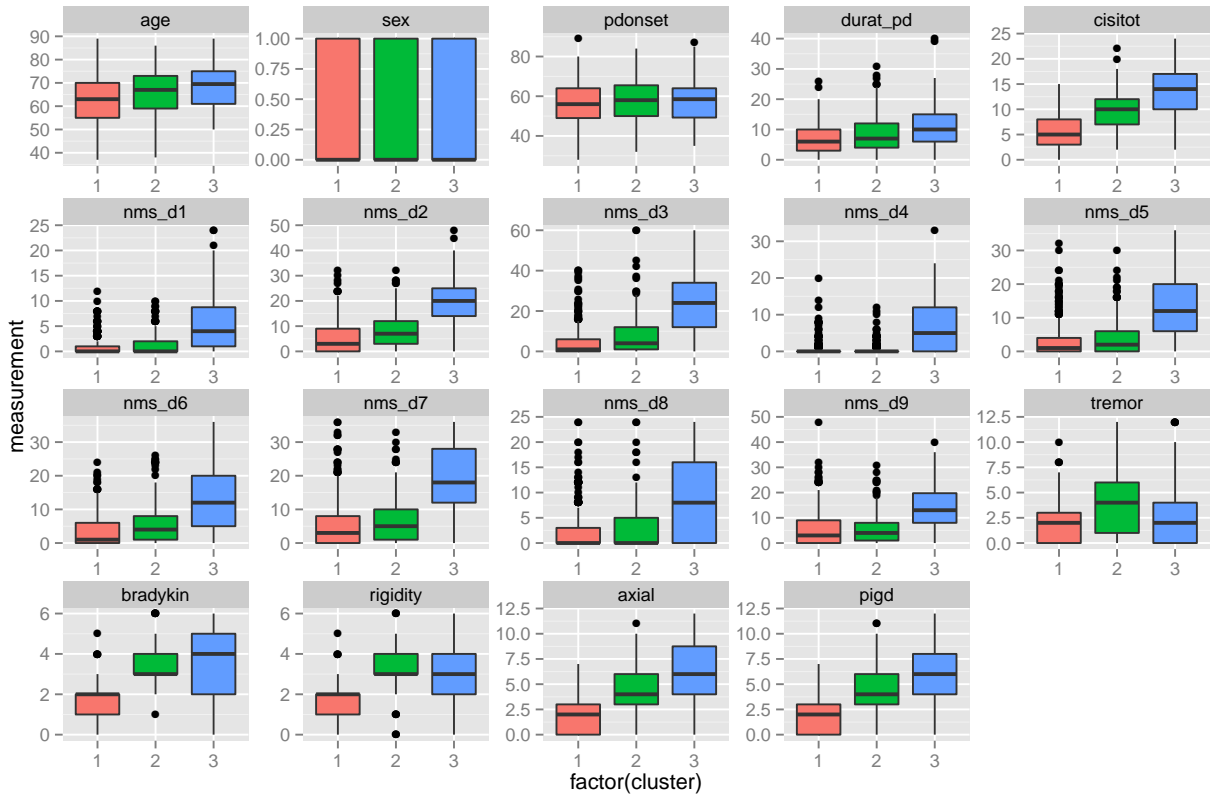


Figure 12: Cluster Summaries,  $k = 3$

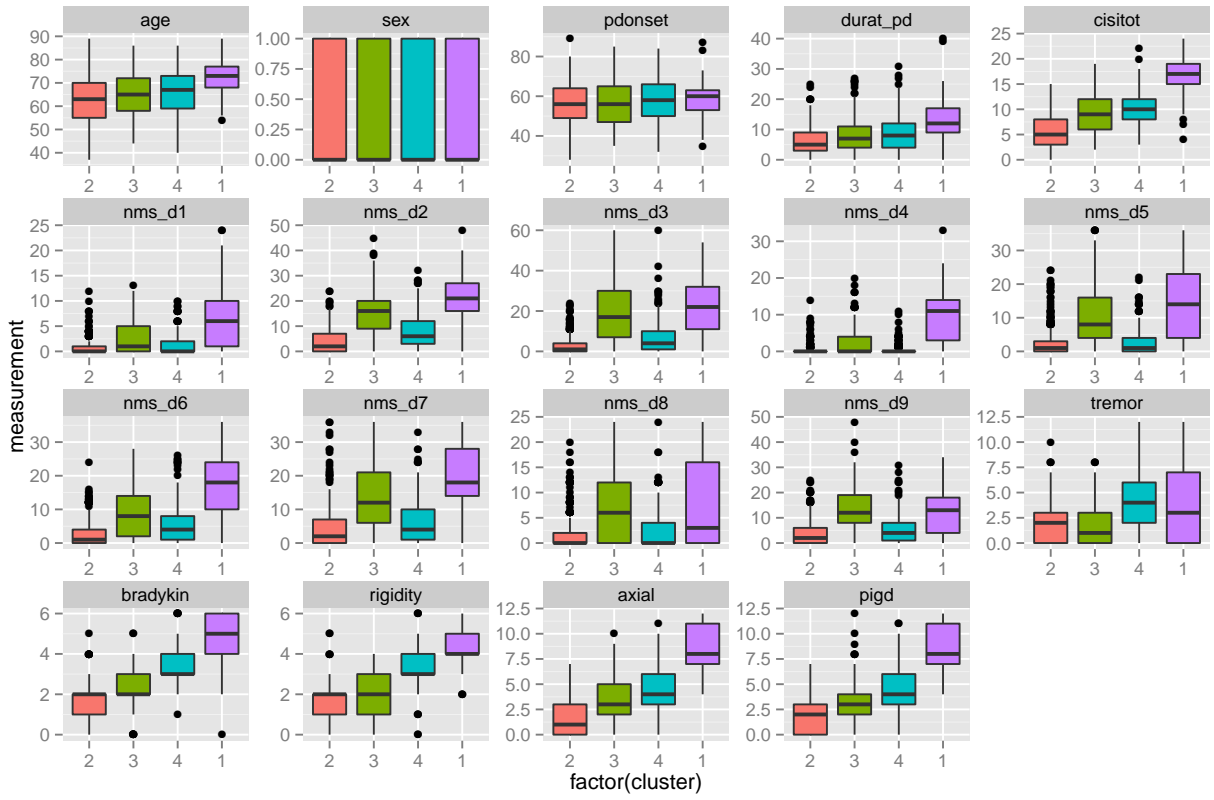


Figure 13: Cluster Summaries,  $k = 4$

## 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.

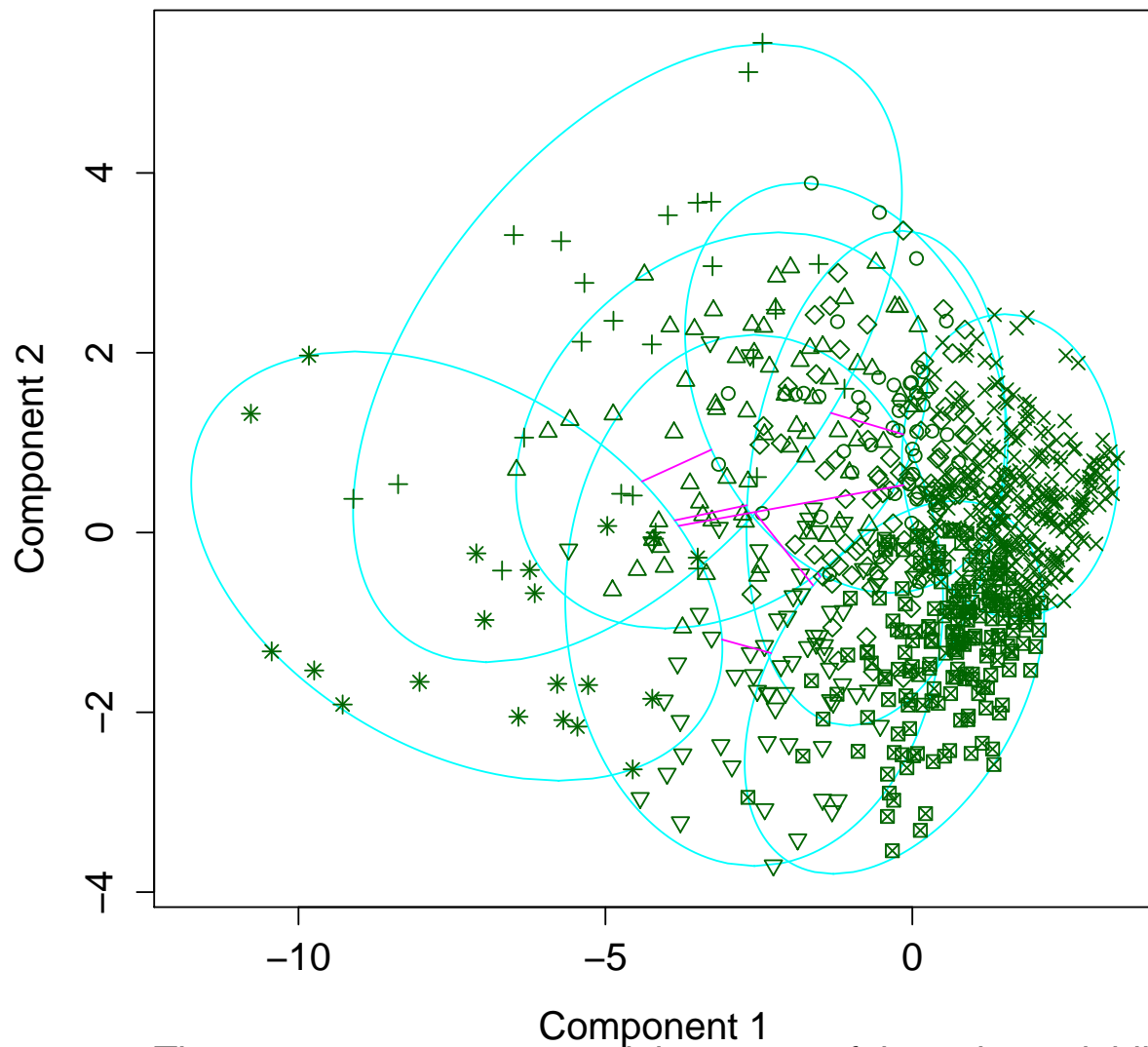
*However*, given that input preferences control how many clusters are found, I don't think it's very useful to have some dozen clusters running around.

#### 3.1.1 Silhouette Plots

Silhouette plot in Figure 14 looks pretty weak, really. Tons of overlap between the clusters.



## AP Silhouette Plot $k = 8$



Component 1  
These two components explain 48.1 % of the point variability.

Figure 14: 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 15

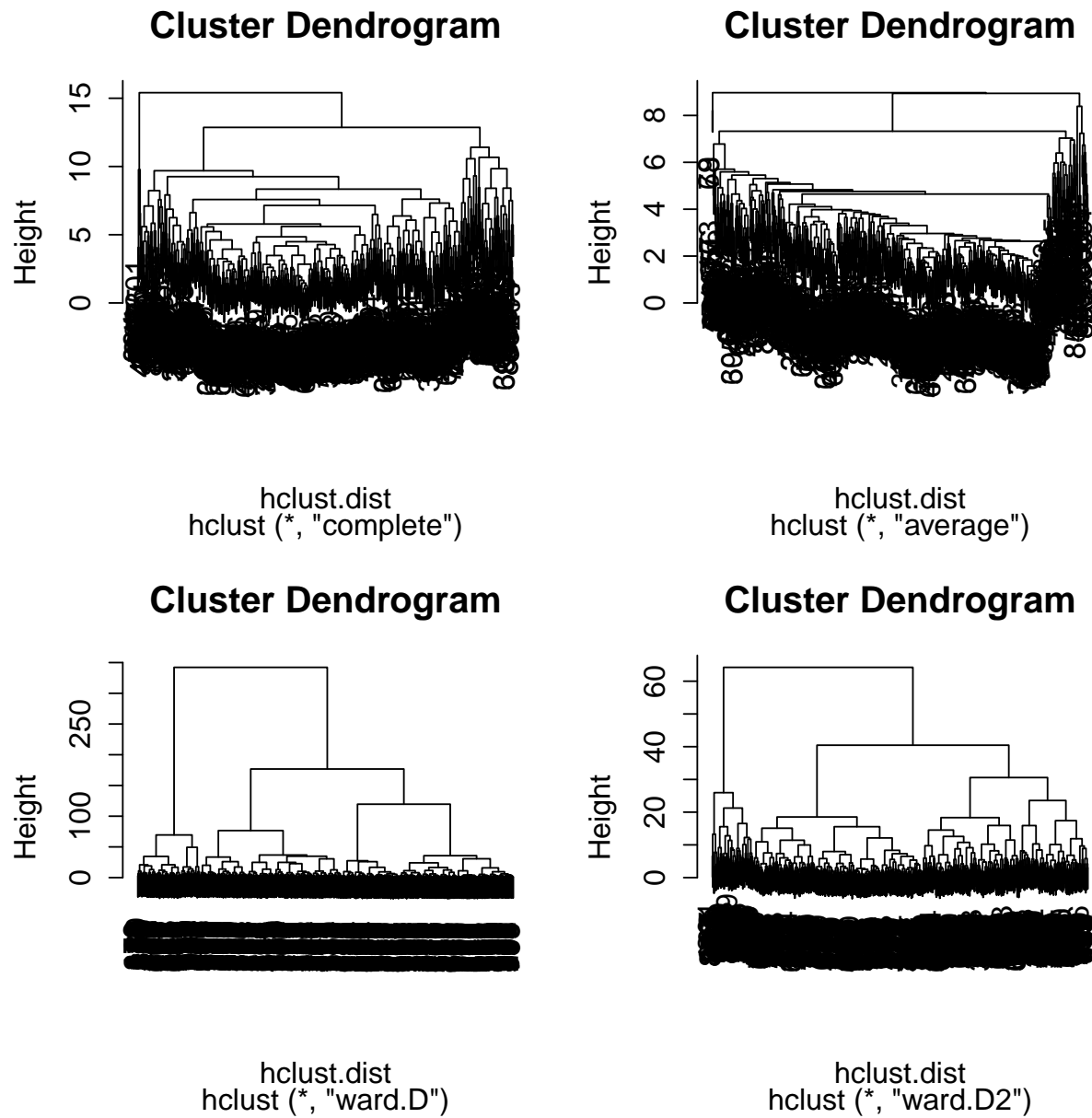


Figure 15: Dendrograms

Method	Condition	n	Figure
Complete	$k = 4$	4 (665/74/35/7)	16
Complete	<code>dynamicTreeCut</code> <sup>5</sup>	11 (7/270/138/83/79/49/46/39/35/35)	17
Ward	$k = 4$	4 (294/236/120/131)	18
Ward	$h = 60$	6 (97/236/120/197/91/40)	19

Table 5: Clusters from Tree Cutting

## 4.2 Cutting Trees

## 4.3 Interpretation

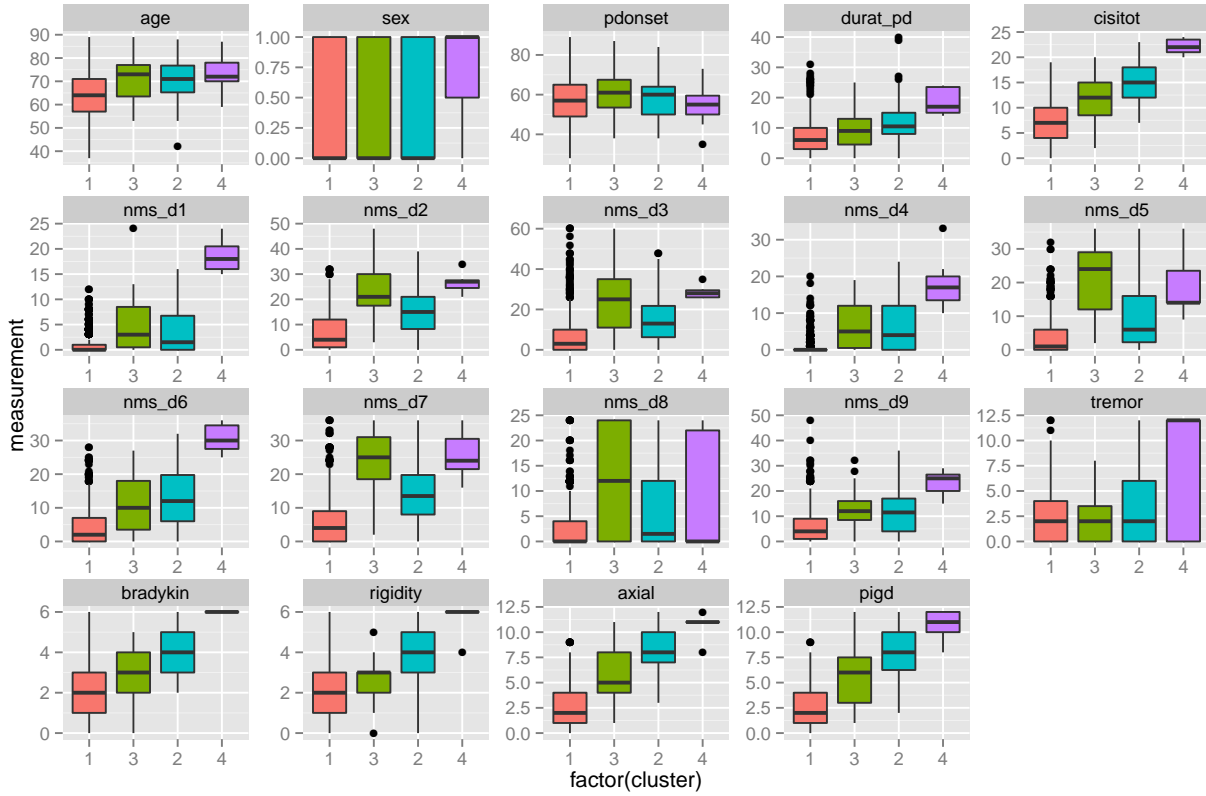


Figure 16: Using maximum (complete linkage) dissimilarity, cutting tree for  $k = 4$

## 4.4 Interpretation

Cluster sizes are available in Table 6

Boxplot summary of clusters available in Figure 20. **Discussion forthcoming.**

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

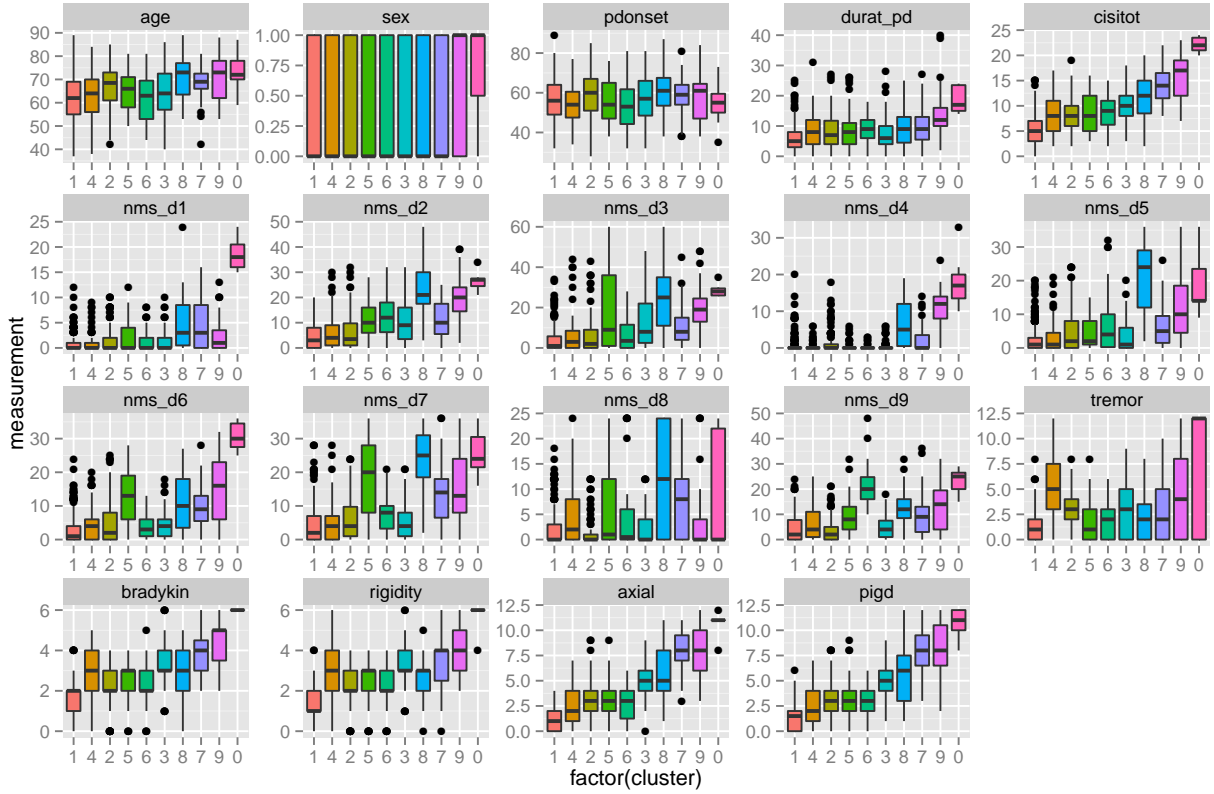


Figure 17: Using maximum (complete linkage) dissimilarity, cutting tree dynamically

## 5 Biclustering

Used BCBimax clustering algorithm. Clusters seem quite sparse.

## 6 Subspace clustering

## 7 Bayesian Networks

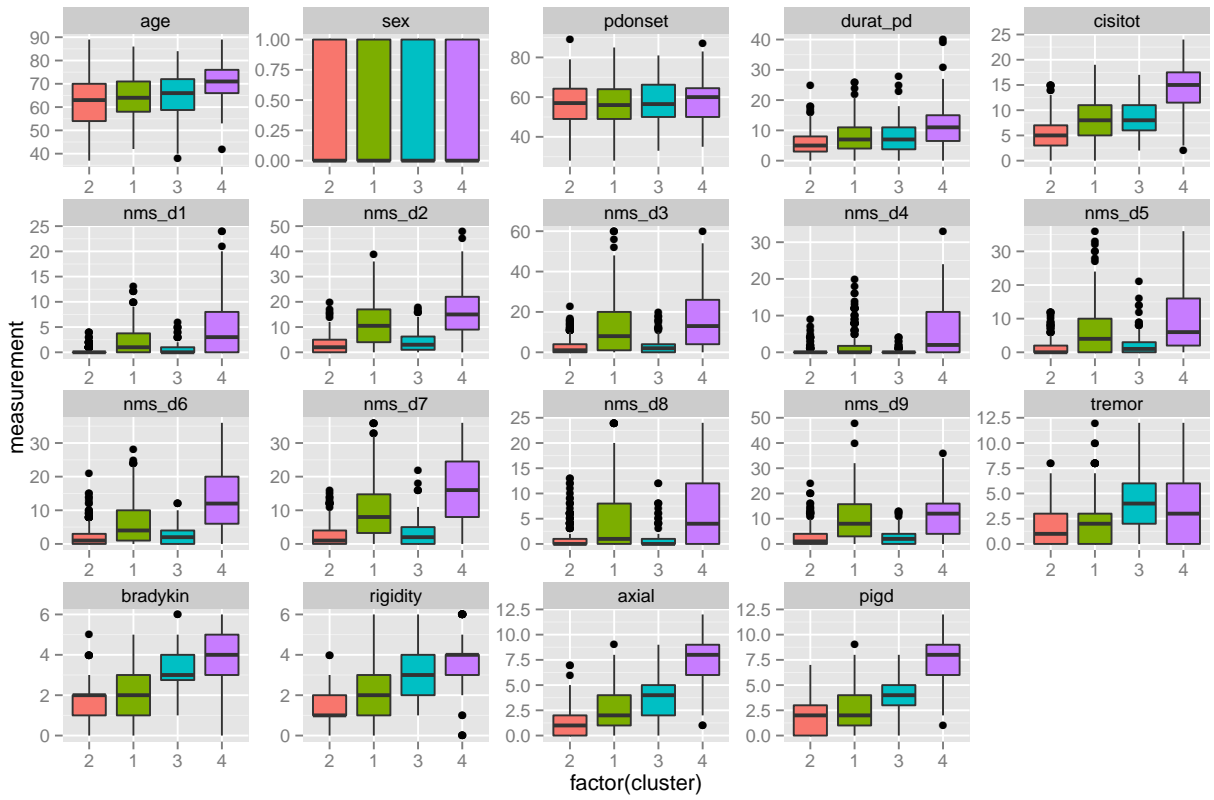


Figure 18: Using Ward (1963) dissimilarity, cutting tree for  $k = 4$

Cluster	Size
1	54
2	68
3	25
4	259
5	102
6	68
7	185
8	20

Table 6: AP Cluster Sizes

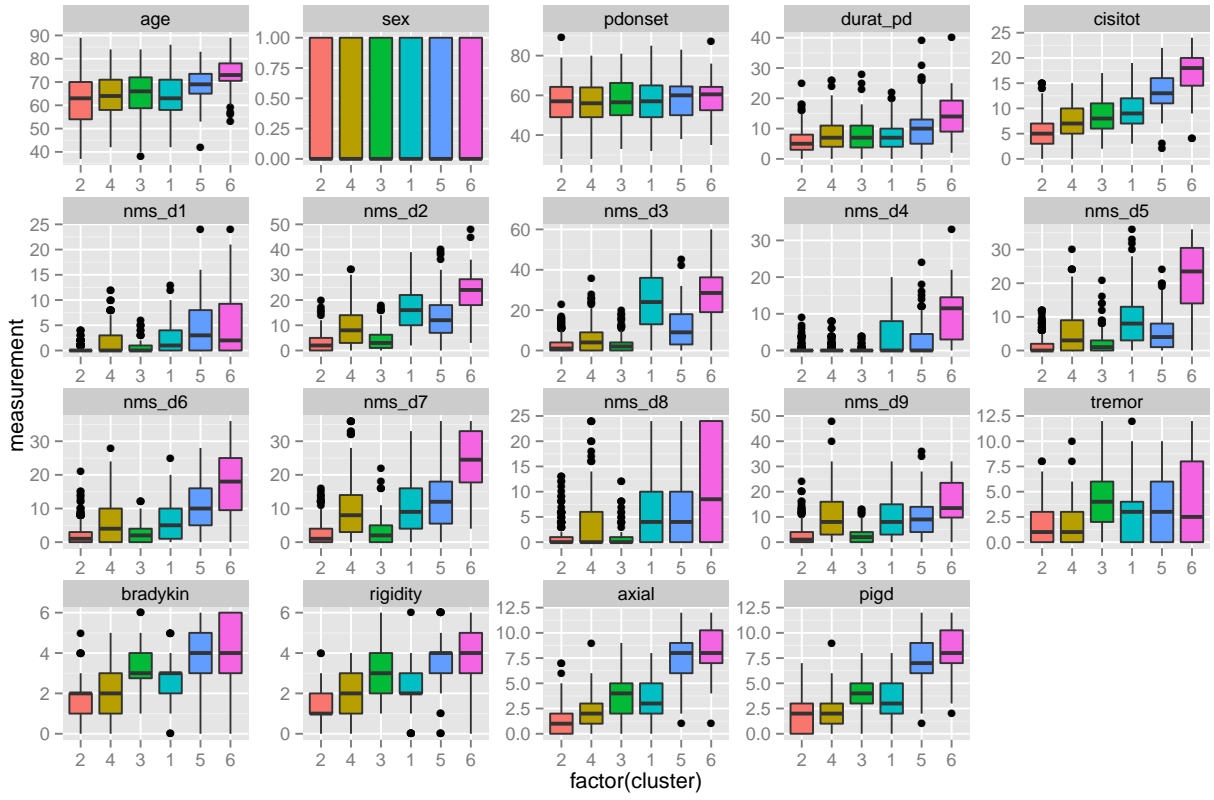


Figure 19: Using Ward (1963) dissimilarity, cutting tree at  $h = 60$

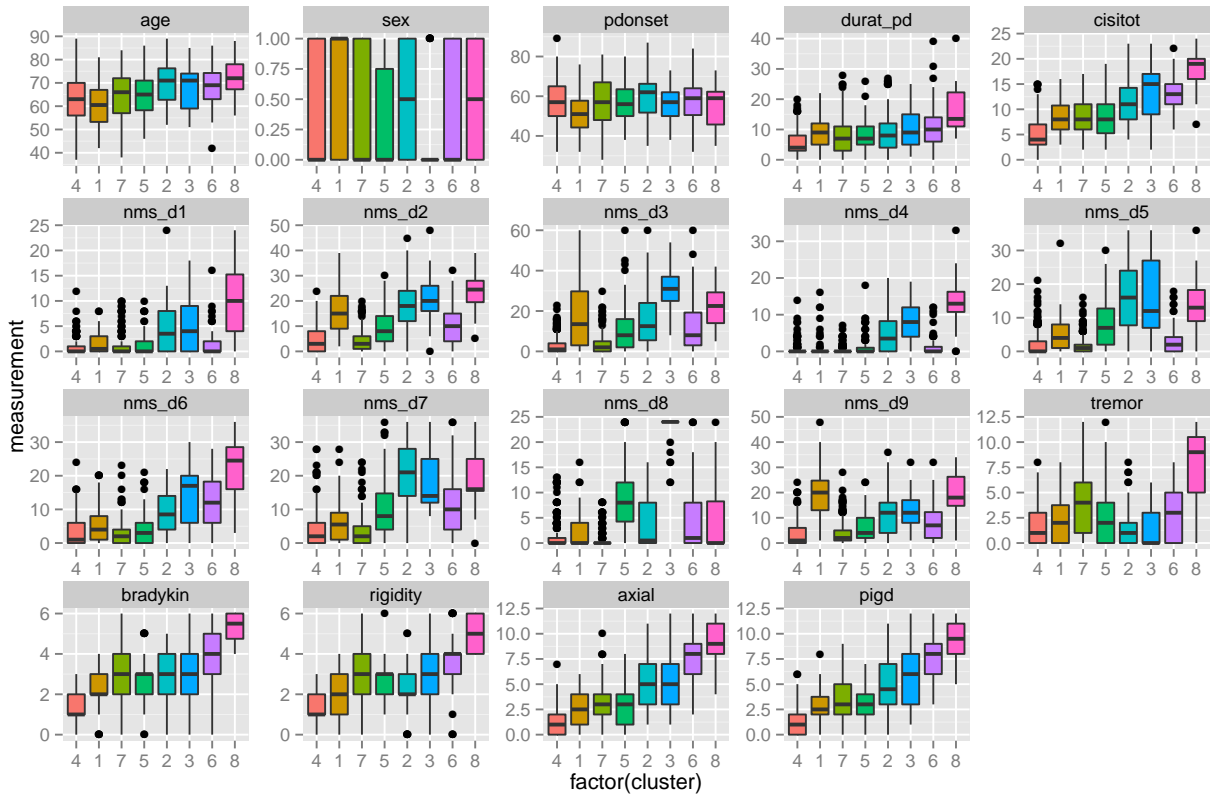


Figure 20: AP Boxplot Summaries

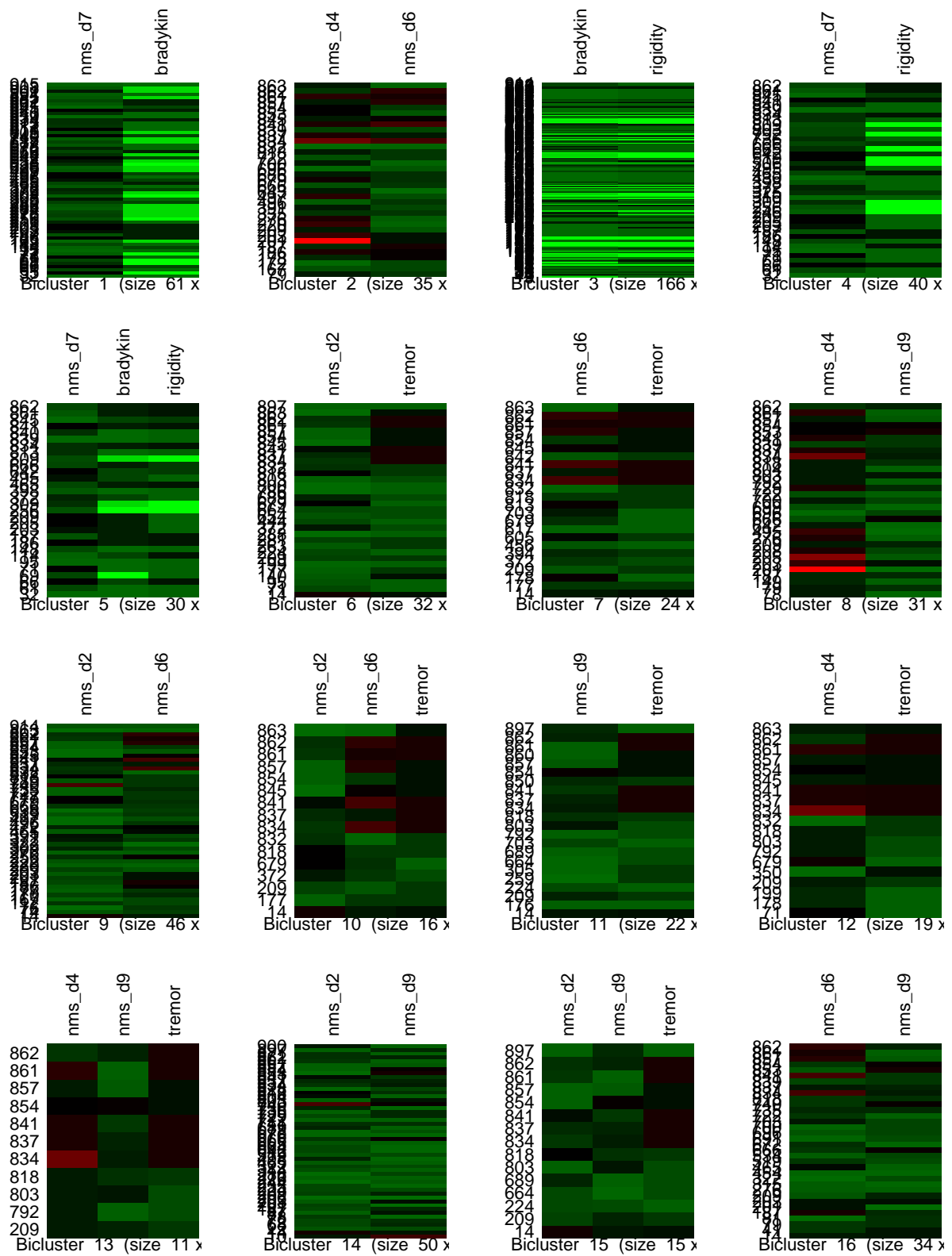


Figure 21: Biclustering  $N = 16$





Figure 22: Bubbleplot  $N = 16$