Cluster Analysis: Identifying Parkinson's Disease Subtypes

Jesse Mu

September 10, 2015

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). It was decided to not impute the data.

1.2 Selected Features

Combination of non-motor scale (NMS) symptoms and standard motor symptoms. PIGD was deleted after 2015-07-16 meeting.

Name	Type	Description
nms_d1	byte	cardiovascular
nms_d2	byte	sleep/fatigue
nms_d3	byte	mood/cognition
nms_d4	byte	percep/hallucinations
nms_d5	byte	attention/memory
nms_d6	byte	gastrointestinal
nms_d7	byte	urinary
nms_d8	byte	sexual function
nms_d9	byte	miscellaneous
tremor	float	tremor
bradykin	float	bradykinesia[1]
rigidity	float	rigidity
axial	float	axial[2]

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

Table 2: Descriptive Statistics

2 Clustering

k-means clustering with $k = 4$ was tried. Statistics for determining the optimal number of clusters were used, but were inconclusive: results in Figure 3. This probably indicates that the data is not very well clustered. $k = 2, 3$ provided models that were too simplistic. $k = 5$ did not provide any new information, but rather just fragmented existing groups.

Criterion	Optimal k
Minimum ASW	2
BIC	18
SSE Scree Plot	Inconclusive
Gap Statistic	4, 13?
Affinity Propagation[3]	8

Table 3: Results of various techniques for determining k

2.1 Decision tree

Decision tree for $k = 4$ created via recursive partitioning is available in Figure 1. More discussion about the decision tree is located in Section 2.2.4.

2.2 Interpretation of Clusters

2.2.1 Cluster summaries

Available in Figure 2. Error bar is standard error.

[3] $\lambda = 0.98$, q = 0, maxits = 1000, convits = 100

2.2.2 Interpretation

k-means clustering ($k = 4$) found four clusters. They are numbered (somewhat confusingly) according to the automatically assigned numbers of the cluster analysis, and not by any intuitive pattern. With a brief description, they are:

4. ($n = 406$) Mildly affected in all domains.

1. ($n = 189$) Severely affected in nonmotor domains; mildly affected in motor domains.

3. ($n = 221$) Severely affected in motor domains; mildly affected in nonmotor domains.

2. ($n = 88$) Severely affected in all domains.

These cluster results are identical to the four clusters found in van Rooden et al [2], which was done with a separate dataset using a different modeling method (expectation-maximization), supporting these subtype classifications. Unlike van Rooden, mean disease durations differences do exist between subtypes 4 (mild) and 2 (severe), likely due to further development of the disease, although the differences between 1 and 3 (nonmotor/motor predominated) subtypes are insignificant, suggesting different developmental paths of the disease.

2.2.3 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.

Post-hoc analysis using Tukey's HSD to examine statistically significant differences between individual means is available in Table 4. Only statistically insignificant relations are provided; all other relations are significant with $p < 0.05$.

2.2.4 Feature importance

Features ranked by information gain with respect to cluster are available in Table 5. Also, in the 4-cluster decision tree in Figure 1, features are ranked implicitly by importance in determining clusters. We see, quite naturally, that standard measures of motor symptoms rank very highly (1, 2, 4, 5) in information gain *except* termor (12). Similarly, bradykinesia (1) is used as the root node of the 4-cluster decision tree, although other motor symptoms are used further down the tree, since immediately successive motor symptom decision nodes would, due to their determination of clusters, be redundant.

The most informative nonmotor symptoms are nms_d2 (sleep/fatigue) at 2, along with nms_d3 (mood/cognition). As discussed later in Section 4.1 these features become critical in one-versus-all decision trees for distinguishing various subtypes. The importance of these nonmotor symptoms confirms the longitudinal study by Fereshtehnejad et al. [1] who cites a 3-cluster PD subtype identification based primarily on nonmotor symptoms including cognitive impairment, rapid eye movement sleep disorder (RBD), anxiety, and depression, conditions that align closely with nms_d2 and nms_d3 as tested in this dataset. More analysis

Variable	Cluster Relation	p
age	3-1	0.724
	4-1	0.428
sex	2-1	0.849
	4-1	0.092
	3-2	0.161
	4-2	0.827
	4-3	0.216
pdonset	2-1	0.147
	3-1	0.370
	4-1	0.859
	3-2	0.803
	4-2	0.305
	4-3	0.700
durat_pd	3-1	0.562
cisitot	3-1	0.523
nms_d1	4-3	0.333
nms_d4	4-3	0.557
nms_d5	4-3	0.856
nms_d8	4-3	0.122
nms_d9	2-1	0.730
	4-3	0.074
tremor	4-1	0.360

Table 4: Tukey's HSD Insignificant Differences

needs to be done on whether there are parallels between Fereshtehnejad's 3-cluster longitudinal study and the clusters found in both this investigation and van Rooden.

Interestingly, demographic information, including durat_pd, age, sex, and pdonset, plays almost no role in the determination of these clusters. That the time of onset of PD or sex nulls important, clinically-relevant questions about the demographic sources of these different subtypes.

2.2.5 Correlation Plots

The interplay between specific symptoms in each of the four clusters was examined in Figure 3, but nothing too interesting was found. There is, perhaps somewhat interestingly, a somewhat higher correlation between bradykinesia, rigidity, and nms_d6 (gastrointestinal), but it was not judged to be significant.

rank	variable	information gain
1	bradykin	0.31574672
2	rigidity	0.29560018
3	nms_d2	0.24218407
4	cisitot	0.22920103
5	axial	0.22780750
6	nms_d3	0.20480570
7	nms_d9	0.15782743
8	nms_d7	0.15290569
9	nms_d5	0.14454931
10	nms_d6	0.14025139
11	nms_d1	0.13212756
12	tremor	0.10937168
13	nms_d4	0.10710526
14	nms_d8	0.10005480
15	durat_pd	0.02876190
16	age	0.02346158
17	sex	0.00000000
18	pdonset	0.00000000

Table 5: Features ranked by information gain

3 Nonmotor-predominant subtype analysis

3.1 k-means sub-subdivision on Cluster 1

In an attempt to understand further the properties of the nonmotor-dominated subtypes, k-means analysis was run again on specifically this subtype to examine any possible patterns.

The same k-determining tests were run on subtype 1 and are displayed in Table 6.

Criterion	Optimal k
Minimum ASW	2
BIC	1 (?)
SSE Scree Plot	Inconclusive
Gap Statistic	3
Affinity Propagation[4]	5

Table 6: Results of various techniques for determining k, applied to subtype 1

Boxplots for k-means run for $k = 2, 3, 4$ can be seen in Figures 4, 5, and 6.

3.2 Interpretation

An interesting set of subtleties occurs when $k = 2$ and 3. When $k = 2$, the two groups are generally somewhat similar, except nms_d2 and nms_d3 move in opposite directions, i.e.

[4] $\lambda = 0.98$, q = 0, maxits = 1000, convits = 100

sub-subtype 2 generally has higher nms_d2 scores but lower nms_d3 scores. In addition, sub-subtype 2 has higher axial scores. This difference can be once again observed in $k = 3$, where subtype 3 is shown to have higher nms_d2 scores, lower nms_d3 scores, and higher axial scores. This may indicate a trend in nonmotor-dominated PD for nms_d2 and nms_d3 to live on opposite ends of a spectrum, i.e. patients especially severe in one may not be as severe in the other. This difference can also been seen in Section 4.1, when OVA decision trees are discussed for the the nonmotor-dominated group.

4 Modeling

One further step of this investigation was to produce accurate, practical models that could be used in a clinical setting to predict the subtype of PD based on previous clustering results. Cluster assignments were treated as labels in a supervised classification problem in an attempt to produce useful models.

4.1 One-versus-all decision trees

While the decision tree in Figure 1 is useful, it could be considered overly complicated. Additionally, a model is not necessarily needed to make simpler diagnoses such as classifying a patient as mildly affected (subtype 4) or severely affected (subtype 2). One-versus-all (OVA) decision trees were thus considered, in order to isolate the classification problem and look at possible distinguishing characteristics of individual subtypes. These OVA decision trees for all 4 subtypes are located in Figures 7, 8, 9, and 10.

4.1.1 4 (mild)

This tree is very odd. The root node considers nms_d9 < 7.5 (miscellaneous) and classifies by far the biggest majority of negative examples when this symptom is milder, even though the subgroup 4 is the mild subtype. After intuitively then classifying on the mildness of rigidity the tree then proceeds to classify once again negative examples based on mildness of nms_d2 (sleep) and positive examples solely based on more severe manifestations of nms_d2, despite the means of these nonmotor symptoms being quite similar according to the boxplots. However, accuracy on the final node is quite poor (61 misclassifications) which could be an indicator that the cluster is not well defined.

4.1.2 1 (nonmotor-predominant)

Quite characteristically, this OVA decision tree clusters first on the severity of bradykinesia. When mild (i.e. < 2.5), all of the subsequent decision nodes ask whether the patient has correspondingly severe forms of primarily nms_d3 (mood/cognition), then nms_d7 (urinary) and nms_d9 (miscellaneous).

When bradykinesia (and all motor symptoms) are relatively severe, however, there is still a small subset of patients who classify into subtype 1 by having severe nms_d2 (sleep) symptoms. While I am not sure if the sample size is large enough here (9/34), I propose this could be one of the "subtypes" of the nonmotor-predominant subtype, i.e. a group that,

when accompanied with high motor symptoms, generally will have high nonmotor symptoms manifest primarily in the form of sleep disorders. This aligns with the findings previously mentioned in Section 3.2 a disjunct between symptoms nms_d2 and nms_d3. As briefly mentioned in that section, axial motor symptom severity also plays a role. In this tree, mildness of axial symptoms is the main classifier of 280 negative examples when identifying nonmotor-predominant subtypes, indicating it is an important distinguishing factor of this unique group of 34 nonmotor-predominant PD patients.

More discussion in the conclusion.

4.1.3 3 (motor-predominant)

This tree classifies overwhelmingly on severity of bradykinesia, with 476 negative examples when bradykinesia is less than 2.5. The resulting tree is quite complex, but generally, nodes check again for severity of motor symptoms (tremor is the next node) and end up classifying positive examples based on both mildness of nonmotor symptoms and severity of motor symptoms. For example, in the furthest right branch, once nms_d2 (as we know, an important feature) is established to be relatively mild (¡ 12), the test for subtype 3 involves two more nodes verifying the severity of rigidty and tremor. Similar behavior can be found in the tree branch testing nms_d7 and nms_d6, then rigidity and axial.

4.1.4 2 (severe)

The OVA tree for severely affected in all areas is predictable, testing entirely on whether or not symptoms (both motor and nonmotor) are relatively severe. Ppositive nodes always appear to the right (no) of less-than checks. Interestingly, however, nms_d4 (percep/hallucinations), previously not of note, is used twice as the root node of a tree and again further down. As the boxplot display in Figure 1 shows, nms_d4 is perhaps the most distinguishing symptom of subtype 4 against nonmotor-predominant subtype 1 in particular, as subtype 1 is relatively mild, in contrast to relatively comparable levels of severity for other nonmotor symptoms. This shows that issues with perception and hallucinations generally occur in only the most severe cases of PD, and are relatively rare when a patient exhibits a nonmotor-predominant form of PD.

5 Conclusion

k-means clustering on this Parkinsons' Disease data set reveals clusters that confirm previous computationally-based findings in the field [2], mainly concerning the identification of four subtypes of Parkinson's disease: mild, nonmotor-predominant, motor-predominant, and severe. The most important nonmotor symptoms in determining these clusters were nms_d2 (sleep) and nms_d3 (mood/cognition), which echo findings of Fereshtehnejad's longitudinal study [1]. More work

Nonmotor symptoms nms_d2 and nms_d3 became critical not only in classification trees distinguishing between the various symptoms but in the nonmotor-predominant subgroup itself, where both standard k-means analysis and decision tree branches show two possible trends in the manifestation of nonmotor-predominant PD:

1. Axial and sleep-severe PD

2. mood/cognition-severe PD

I am, however, not sure if these groups hae enough members to warrant a subtype, or whether this could just be chance or noise.

It remains to be seen whether these classification models, especially the one-vs-all decision trees, are useful in clinical practice.

5.1 Bayesian Networks

5.1.1 On all data

I decided to discretize the data into three uniform-width groups based on the scales of each symptom. In other words, each symptom was discretized into a mild, moderate, and severe bin. Continuous data was unreliable on my computer, and updating intricately connected nodes like nms_d2 resulted in slowdowns and crashes on my computer.

Two bayesian network algorithms were tried: the default Bayesian score-search algorithm and the PC conditional independence tests algorithm. I couldn't find the exact name of the Bayesian search implementation, but it was the default method used by GeNIe. GeNIe files will be attached electronically.

I assume these models are to be looked at by Dr. Martín. I have not done too much investigation myself, as I'm not exactly sure what I'm looking for).

5.1.2 On nms-dominated data

I tried to construct Bayesian networks based on the nms-dominated subtype, but the data was too sparse to create a very informative network, even when leaving the information continuous. However, I'm not sure this is necessary. If it is, I can work on this problem more.

References

[1] Fereshtehnejad et al (June 15, 2015). New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression

[2] van Rooden et al (2010). The Identification of Parkinson's Disease Subtypes Using Cluster Analysis: A Systematic Review

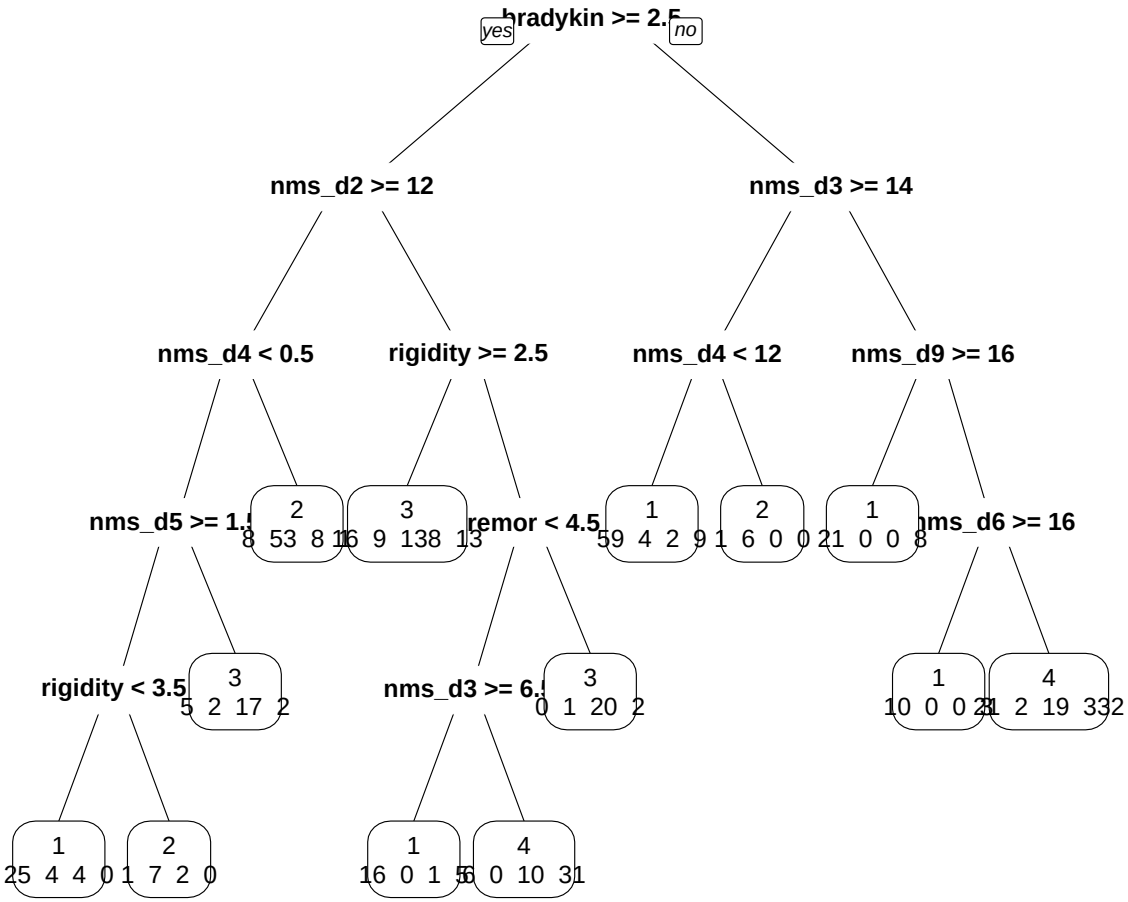

Figure 1: Decision Tree from k-means clustering, 4 clusters

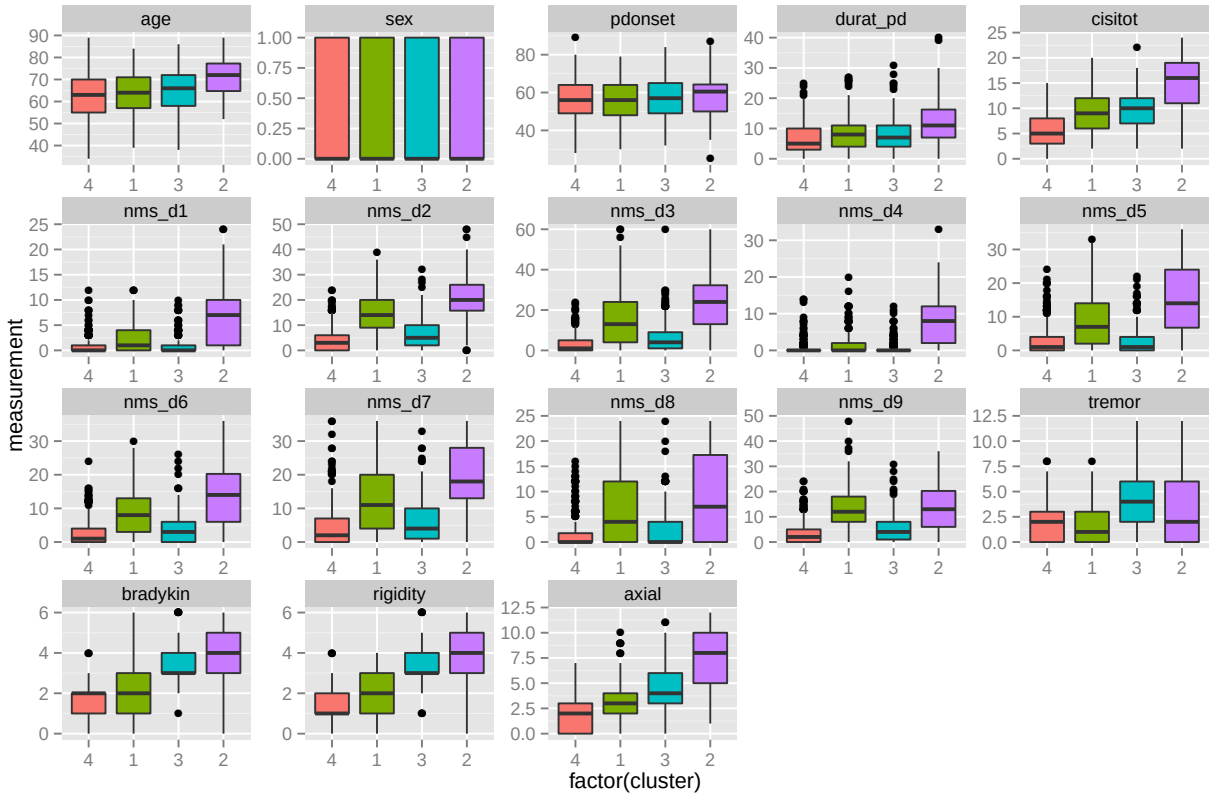

Figure 2: Cluster Summaries, $k = 4$

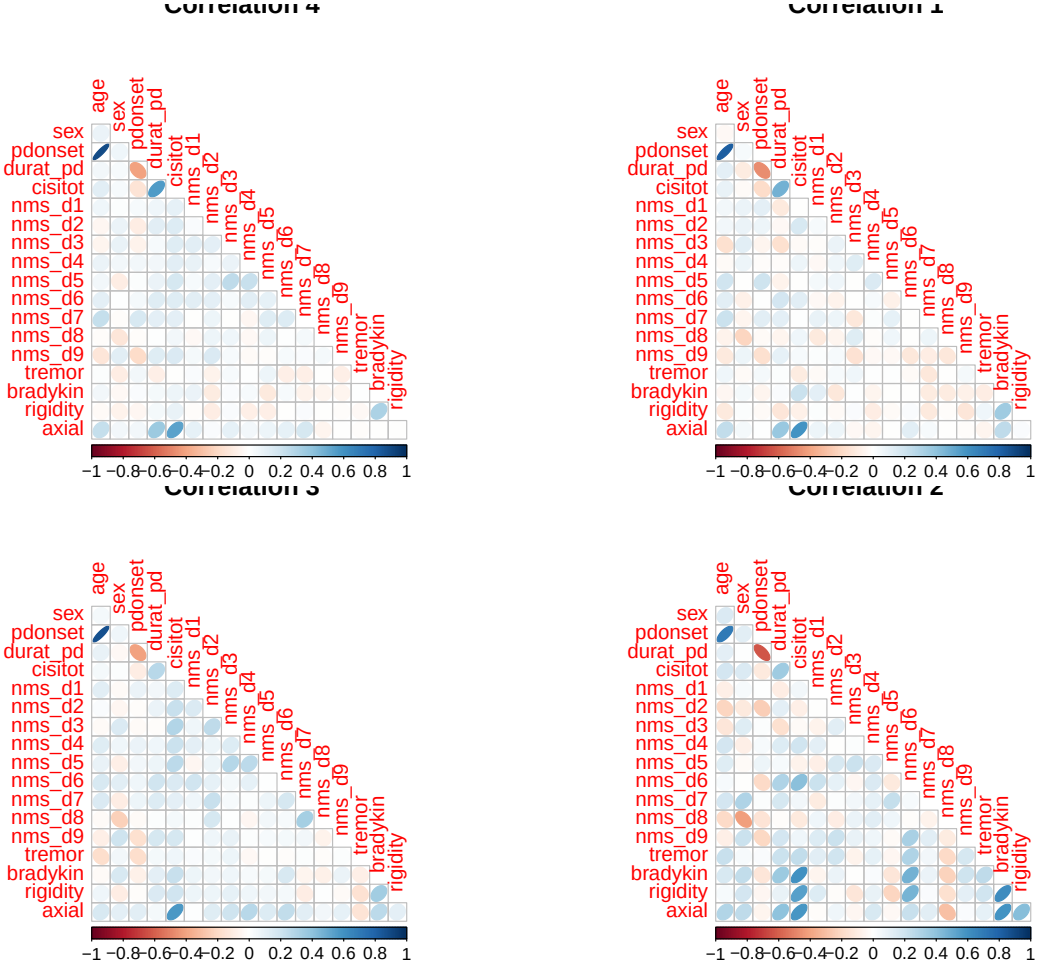

Figure 3: Correlation plots

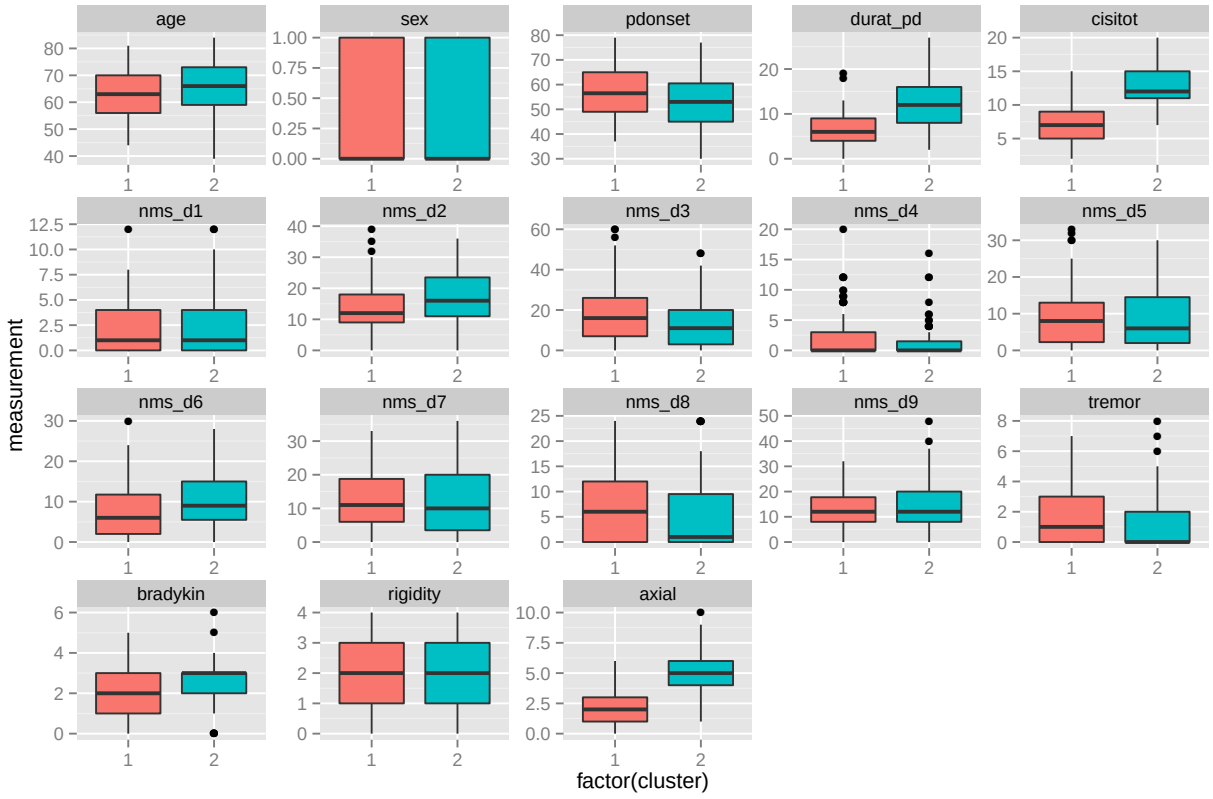

Figure 4: Clustering on nonmotor group: $k = 2$

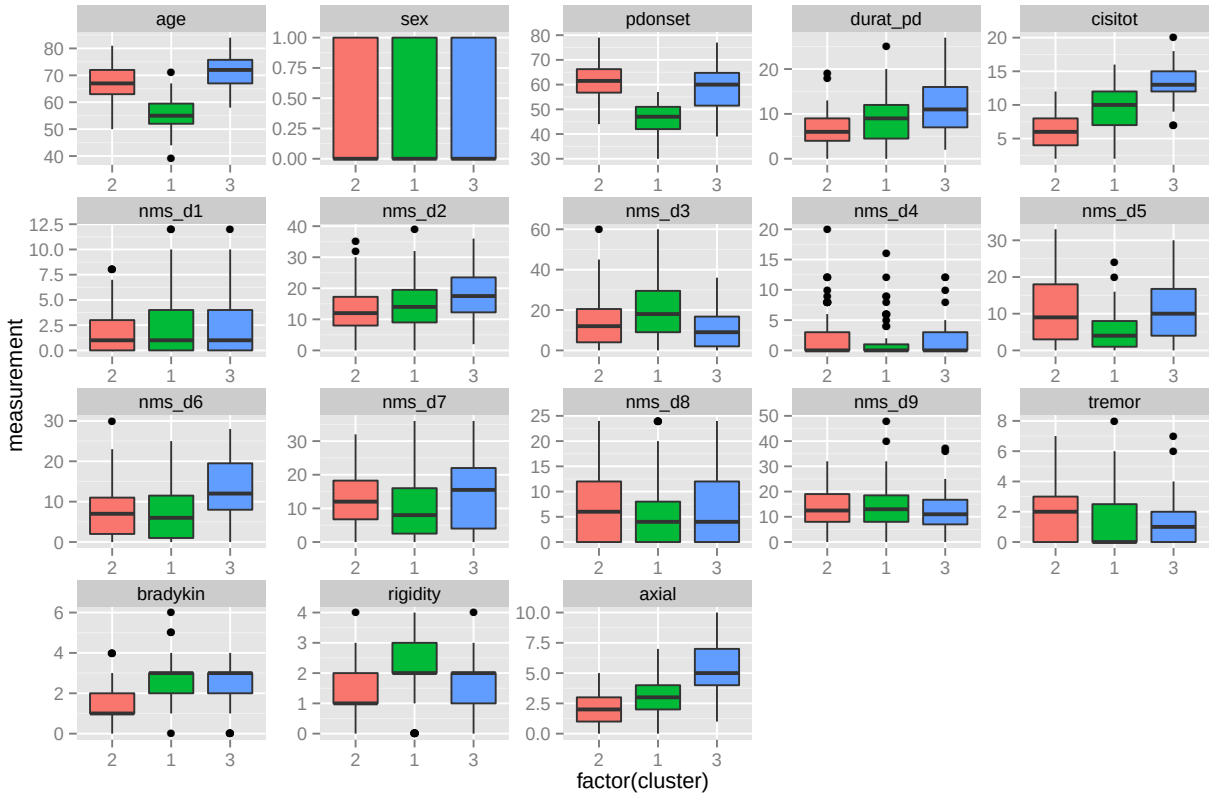

Figure 5: Clustering on nonmotor group: $k = 3$

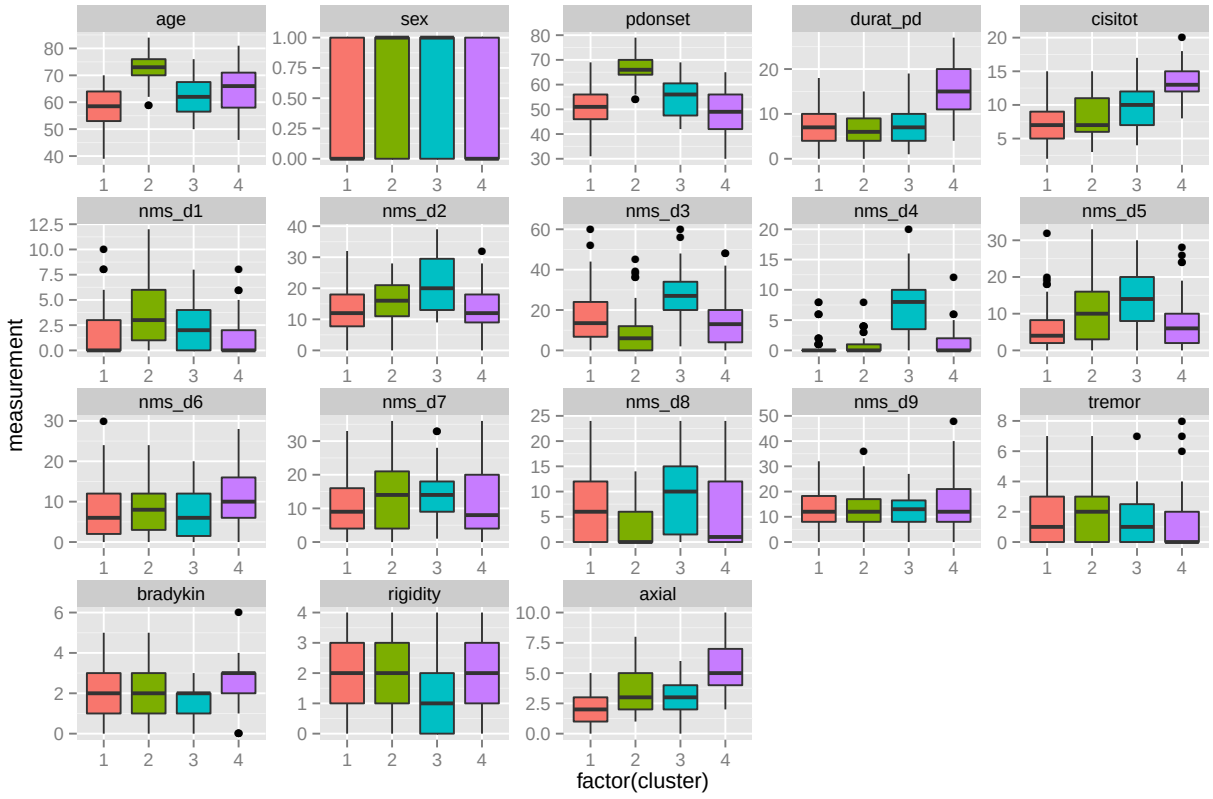

Figure 6: Clustering on nonmotor group: $k = 4$

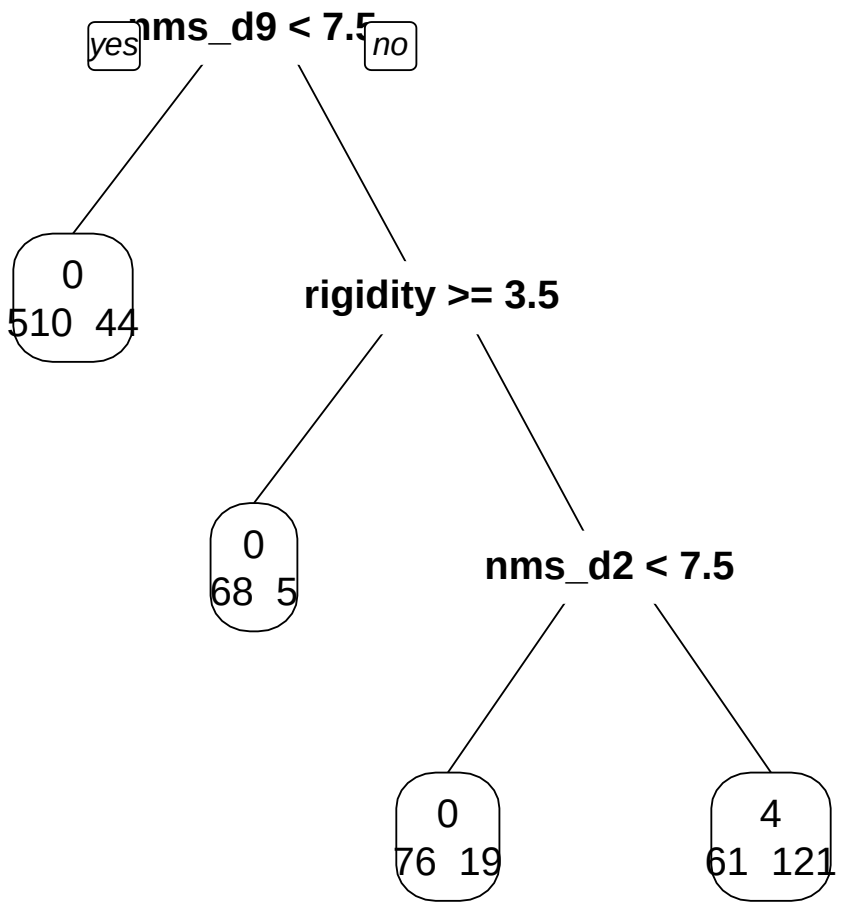

Figure 7: Cluster 4 (mild) vs all

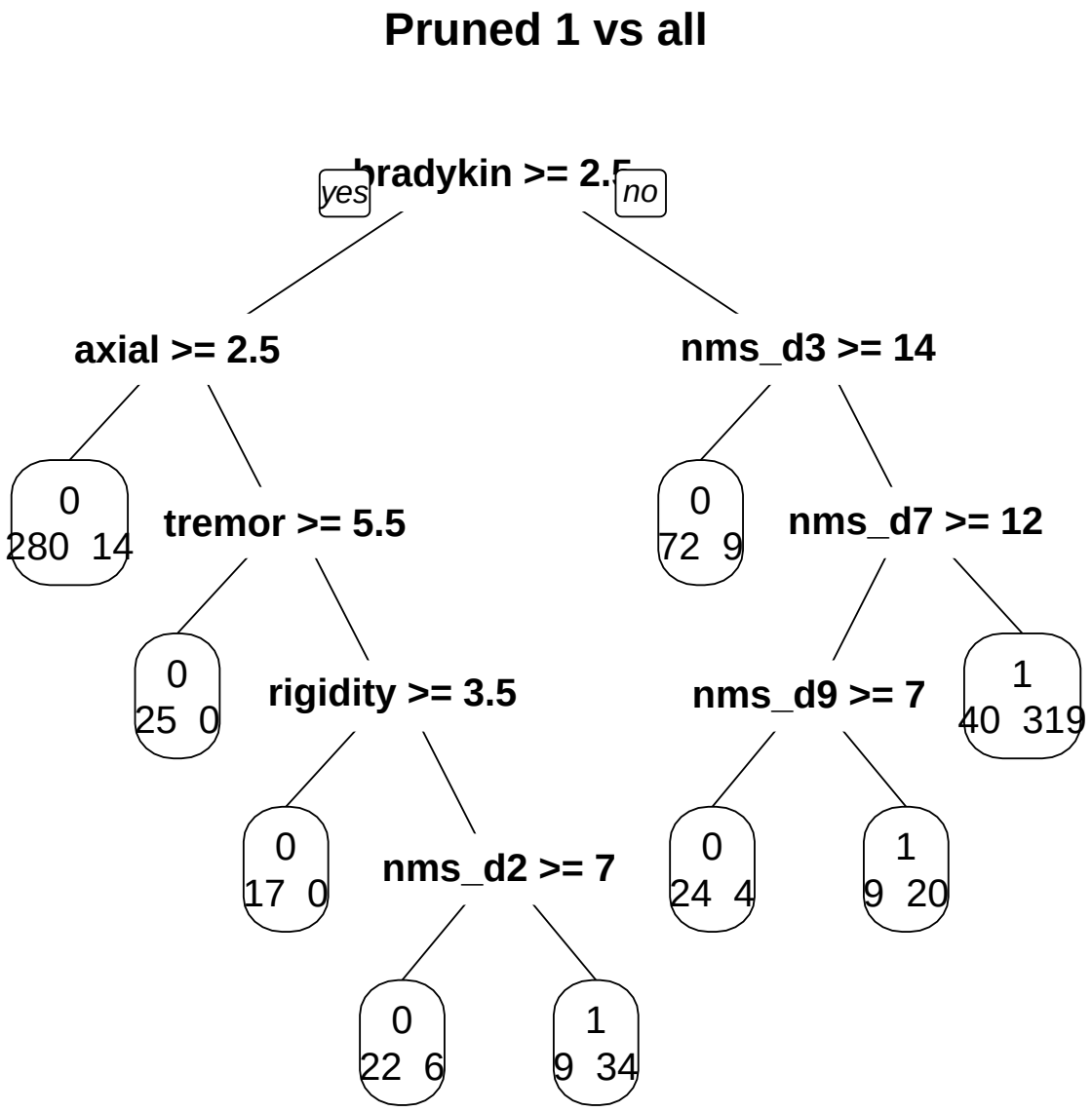

Figure 8: Cluster 1 (nonmotor-dominated) vs all

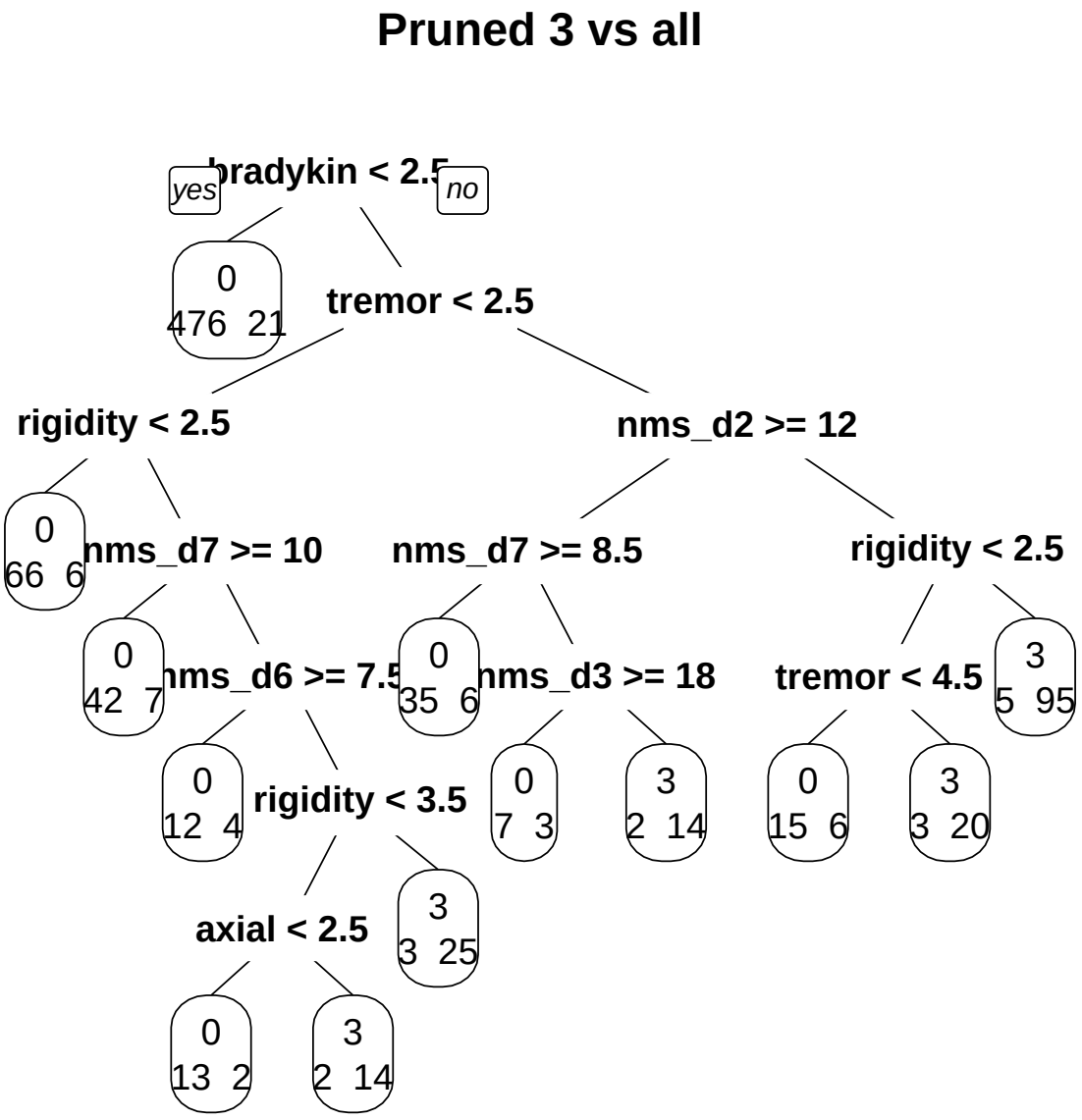

Figure 9: Cluster 3 (motor-dominated) vs all

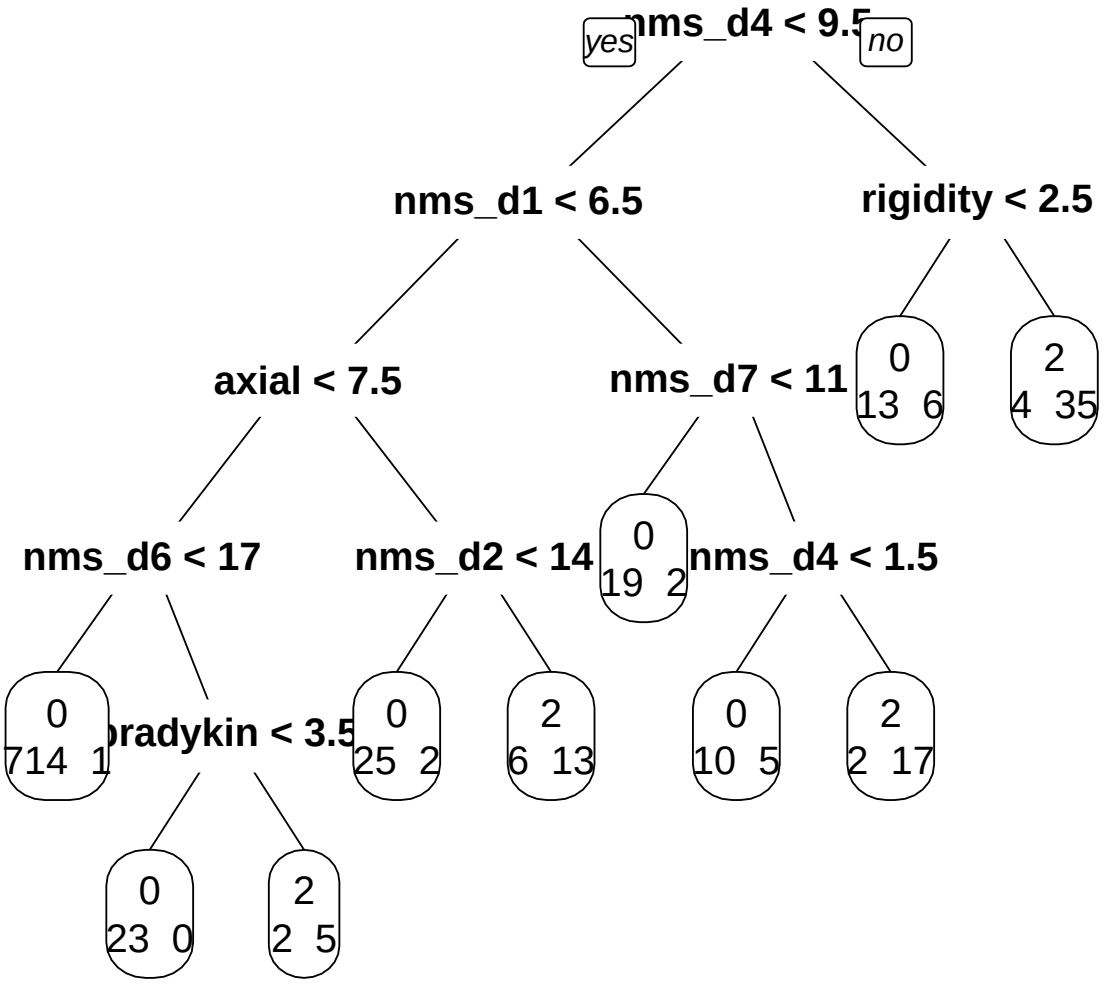

Figure 10: Cluster 2 (severe) vs all