Cluster Analysis: Identifying Parkinson's Disease Subtypes

Jesse Mu

March 3, 2016

Contents

1 New Content

Section 6 on longitudinal analysis; section 7 on Extended Nonmotor Symptoms; Section 8.3 on new conclusions from these sections.

2 Preprocessing

2.1 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. Data was scaled to $\mu = 0, \sigma = 1$ during clustering and modeling, then unscaled for visualization.

2.2 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

3 Clustering

k-means clustering with $k = 4$ was tried. Statistics for determining the optimal number of clusters were used, but were generally 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
Affinity Propagation	8
clValid stability measures	4

Table 3: Results of various techniques for determining k

3.1 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 3.2.4.

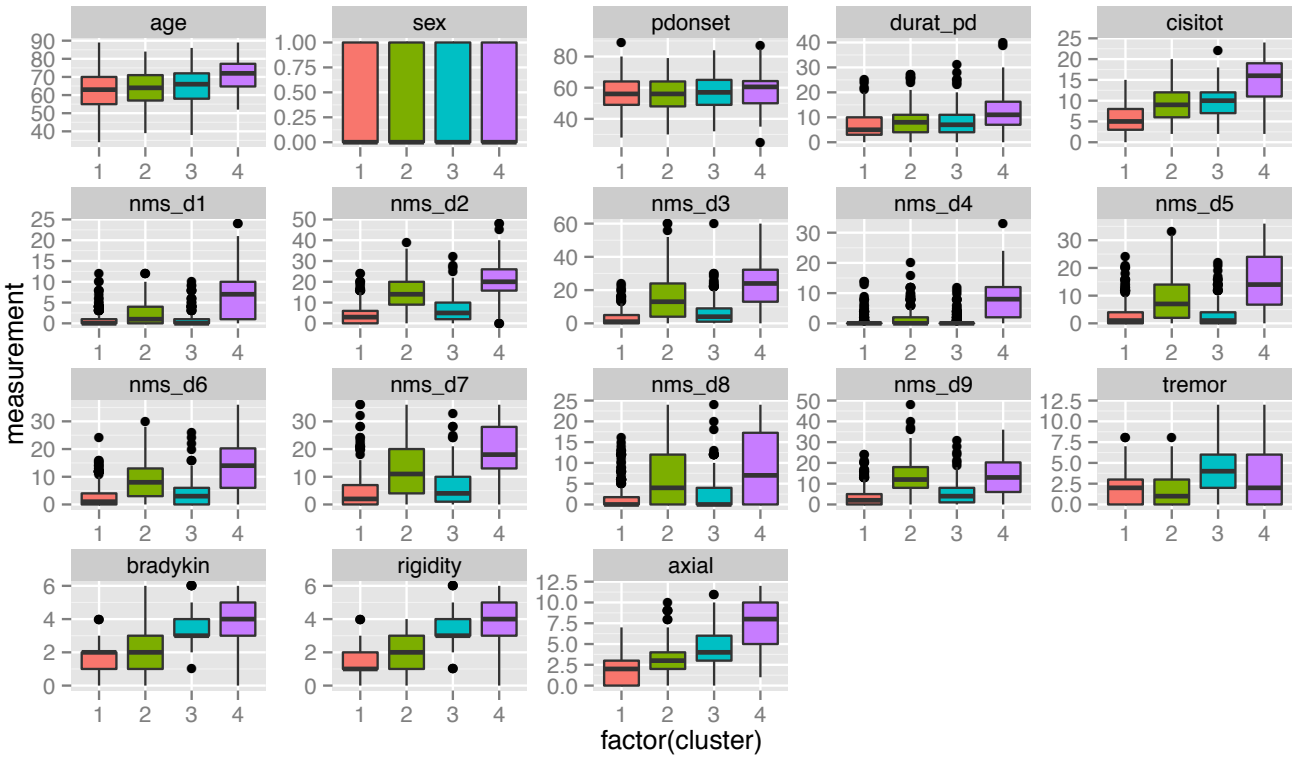


Figure 2: Cluster Summaries, $k = 4$

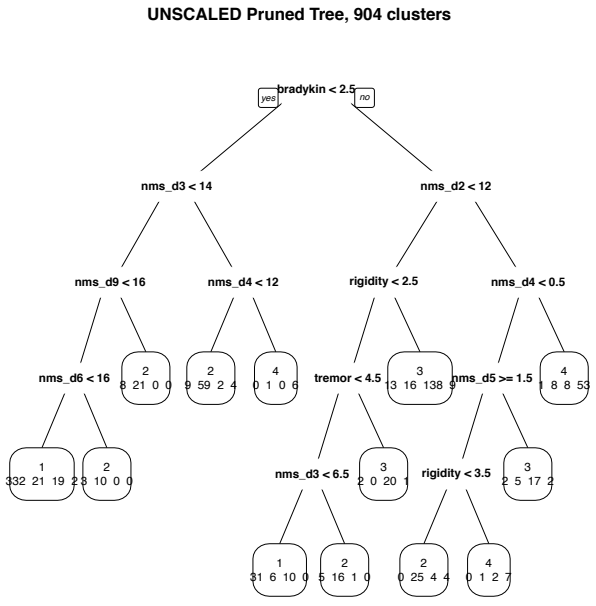


Figure 1: Decision Tree from k-means clustering, 4 clusters

3.2 Interpretation of Clusters

3.2.1 Cluster summaries

Available in Figure 2. Error bar is standard error.

3.2.2 Interpretation

k-means clustering ($k = 4$) found four clusters. With a brief description, they are:

1. ($n = 406$) Mildly affected in all domains.
2. ($n = 189$) Severely affected in nonmotor domains; mildly affected in motor domains.
3. ($n = 221$) Severely affected in motor domains; mildly affected in nonmotor domains.
4. ($n = 88$) Severely affected in all domains.

3.2.3 Statistical Significance Tests, $k = 4$

For each variable i and cluster means $\mu_i^1, \mu_i^2, \mu_i^3, \mu_i^4$, we use one-way ANOVA for multiple means and reject the null hypothesis that $\mu_i^1 = \mu_i^2 = \mu_i^3 = \mu_i^4$ 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. For brevity, only statistically insignificant relations are provided; all other relations are significant with $p < 0.05$.

Variable	Cluster Relation	p
age	2-1	0.428
	3-2	0.724
sex	2-1	0.0918
	3-1	0.216
	4-1	0.827
	4-2	0.849
	4-3	0.161
pdonset	2-1	0.859
	3-1	0.700
	4-1	0.305
	3-2	0.370
	4-2	0.147
	4-3	0.803
durat_pd	3-2	0.562
cisitot	3-2	0.522
nms_d1	3-1	0.333
nms_d4	3-1	0.557
nms_d5	3-1	0.856
nms_d8	3-1	0.122
nms_d9	3-1	0.0735
	4-2	0.730
tremor	2-1	0.360

Table 4: Tukey's HSD Insignificant Differences

3.2.4 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 (ranks 1, 2, 4, 5) in information gain *except* tremor (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 5.1 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 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 is largely irrelevant provides an important negative answer to clinically-relevant questions about the demographic sources of these different subtypes.

rank	variable	information gain
1	bradykin	0.316
2	rigidity	0.296
3	nms_d2	0.242
4	cisitot	0.229
5	axial	0.228
6	nms_d3	0.205
7	nms_d9	0.158
8	nms_d7	0.153
9	nms_d5	0.145
10	nms_d6	0.140
11	nms_d1	0.132
12	tremor	0.109
13	nms_d4	0.107
14	nms_d8	0.100
15	durat_pd	0.0288
16	age	0.0235
17	sex	0.000
18	pdonset	0.000

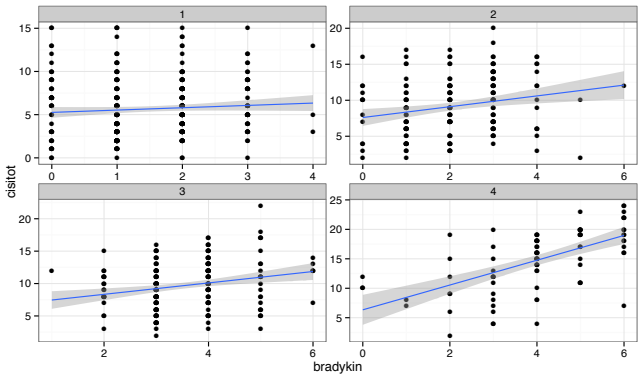
Table 5: Features ranked by information gain

3.2.5 Correlation Plots

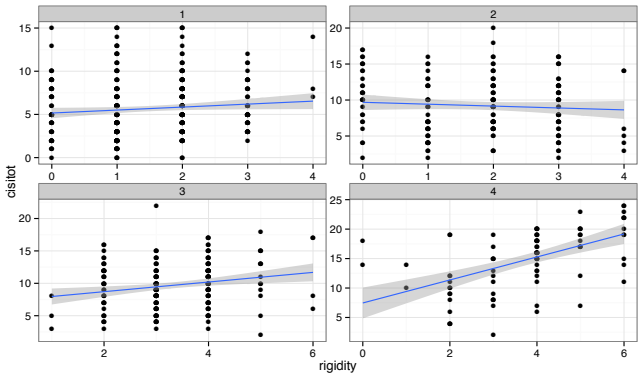
The interplay between specific symptoms in each of the four clusters was examined in Figure 5. There are two points of note. The first is that there is a higher correlation in cluster 4 (severe) between overall severity (cisitot) and bradykinesia and rigidity, illustrated in Figure 3. Second, there exists a somewhat higher correlation between bradykinesia, rigidity, and nms_d6 (gastrointestinal) in cluster 4, illustrated in Figure 4. These differences are statistically significant; correlation tests are located in Table 6. I am unsure of the significance or proper interpretation of these results.

Cluster	Variables	95% CI	p
1	bradykin, cisitot	[-0.0225, 0.171]	0.131
	rigidity, cisitot	[-0.000406, 0.192]	0.0510
	bradykin, nms_d6	[-0.0634, 0.131]	0.493
	rigidity, nms_d6	[-0.101, 0.0932]	0.934
2	bradykin, cisitot	[0.0786, 0.351]	$0.00248^{(**)}$
	rigidity, cisitot	[-0.215, 0.069]	0.310
	bradykin, nms_d6	[-0.152, 0.133]	0.897
	rigidity, nms_d6	[-0.123, 0.163]	0.781
3	bradykin, cisitot	[0.0995, 0.350]	$0.000620^{(***)}$
	rigidity, cisitot	[0.0687, 0.322]	$0.00298^{(**)}$
	bradykin, nms_d6	[0.0350, 0.292]	$0.0134^{(*)}$
	rigidity, nms_d6	[-0.0846, 0.179]	0.478
4	bradykin, cisitot	[0.454, 0.724]	$3.97 \times 10^{-10(***)}$
	rigidity, cisitot	[0.375, 0.675]	$4.99 \times 10^{-08(***)}$
	bradykin, nms_d6	[0.297, 0.624]	$2.60 \times 10^{-06(***)}$
	rigidity, nms_d6	[0.278, 0.611]	$6.43 \times 10^{-06(***)}$

Table 6: Correlation tests. (*) $p < 0.05$, (**) $p < 0.01$, (***), $p < 0.001$

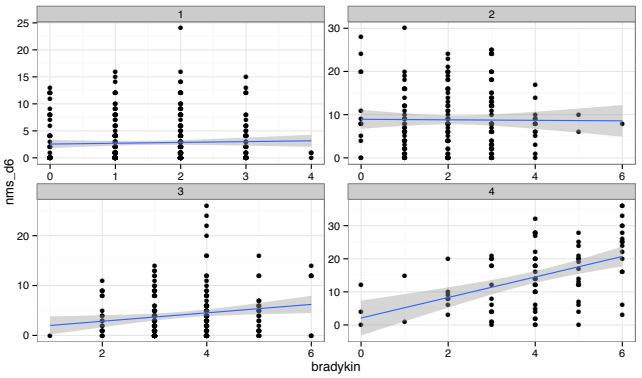


(a)

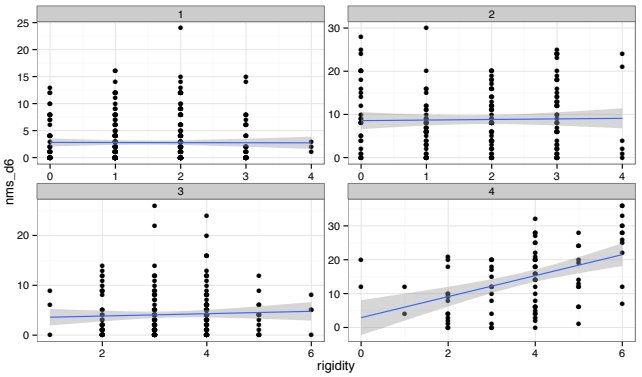


(b)

Figure 3: Relationship between (a) bradykinesia, (b) rigidity and overall severity (cisitot). Shaded band is 95% confidence interval.



(a)



(b)

Figure 4: Relationship between (a) bradykinesia, (b) rigidity and nms_d6 (gastrointestinal). Shaded band is 95% confidence interval.

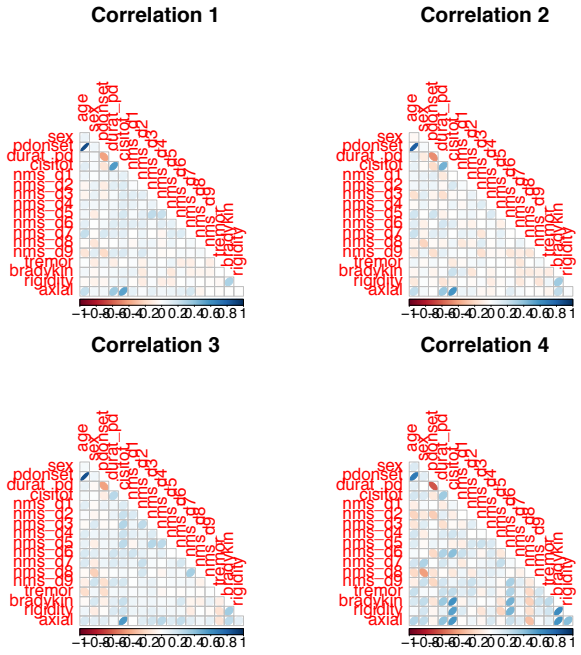


Figure 5: Correlation plots

4 Nonmotor-predominant subtype analysis

4.1 k-means sub-subdivision on Cluster 2

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 2 and are displayed in Table 7.

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

Table 7: Results of various techniques for determining k, applied to subtype 2

Boxplots for k-means run for $k = 2, 3, 4$ can be seen in Figures 6, 7, and 8. Clusters are ordered by increasing cisitot.

4.2 Interpretation

An interesting set of subtleties occurs when $k = 2$ and 3. When $k = 2$, the two groups are divided generally by PD severity (see cisitot and especially axial). The specific symptoms of the two groups follow this trend, except nms_d3 and tremor, which are actually decreasing, and other symptoms like rigidity, nms_d4, and nms_d9, which are more indeterminate.

When $k = 3$, the symptoms that continue show a nonmonotonically increasing trend are nms_d2, tremor, and rigidity scores, where patients in the 3rd subtype exhibit lower severities. nms_d4 and nms_d9 differences turn out to be not as pronounced.

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

5 Further 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 obtained from previous k-means investigation were treated as labels in a supervised classification problem in an attempt to produce useful and easily interpretable models.

5.1 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 1) or severely affected (subtype 4). 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 9, 10, 11, and 12. Trees are pruned by selecting the version of tree with the minimum 10-fold cross-validated error.

5.1.1 1 (mild)

The tree for the mild subtype classifies mainly based on negative responses to nodes asking whether the patient has a relatively severe manifestation of a symptom. The majority of examples are classified by following the bradykinesia < 2.5, which subsequently tests the severity of several nonmotor symptoms. Most of subtype 1 patients that score relatively mildly on these scales are classified this way. There are also small populations of patients who 1) score higher on bradykinesia but lower with axial, tremor, rigidity, and nms_d2 and 2) score higher in nms_d2 (sleep) but lower with nms_d7 (urinary).

5.1.2 2 (nonmotor-predominant)

The decision tree for the nonmotor-predominant subtype is quite simple. Interestingly, although nms_d9 (miscellaneous) is not the most important nonmotor symptom, since the information gain is less than nms_d2 and nms_d3

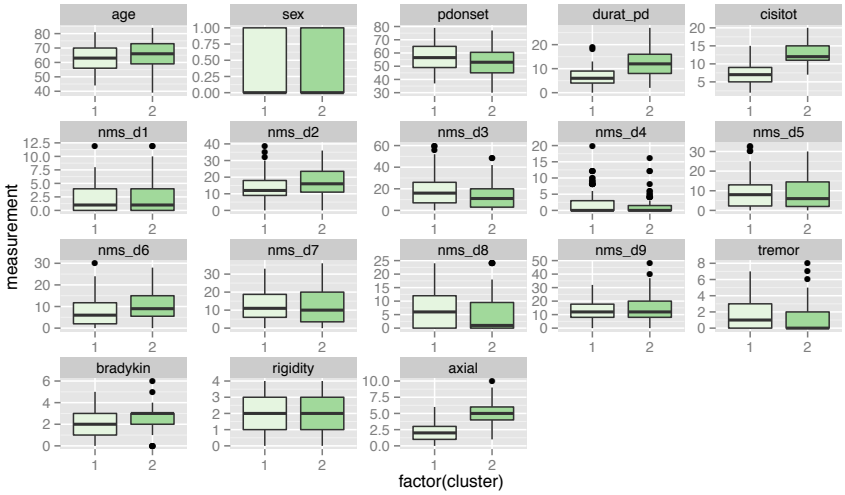


Figure 6: Clustering on nonmotor group: $k = 2$

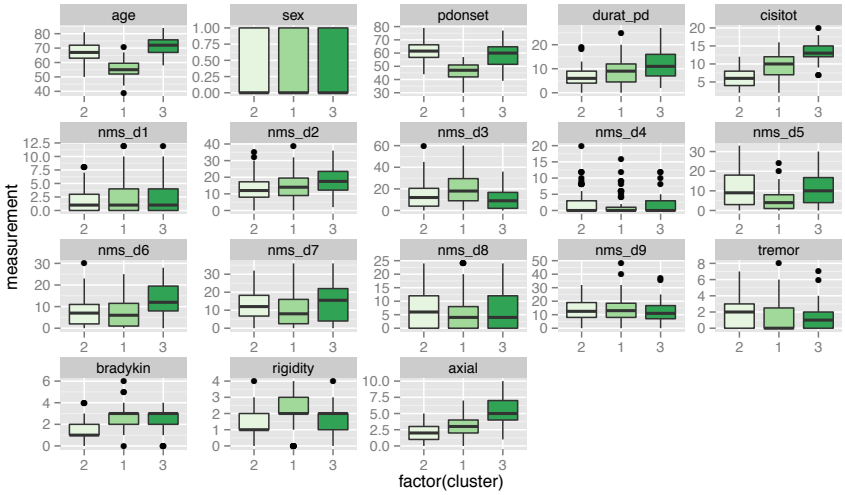


Figure 7: Clustering on nonmotor group: $k = 3$

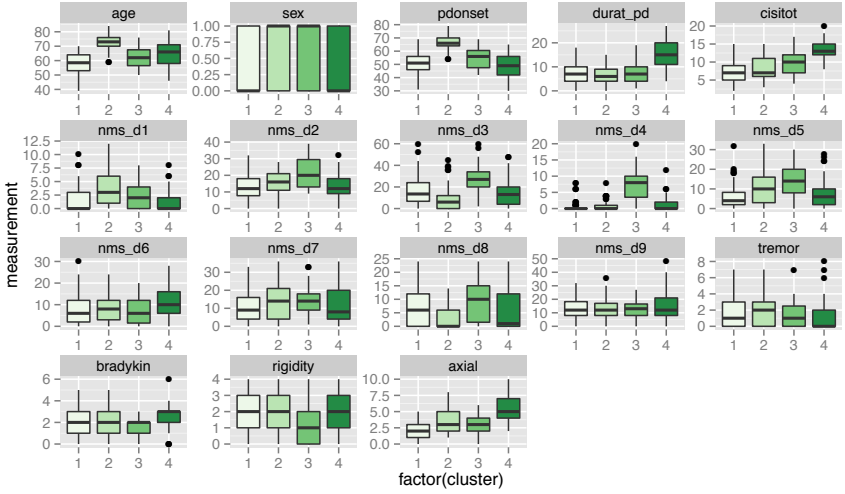


Figure 8: Clustering on nonmotor group: $k = 4$

and it does not appear very high in the 4-class decision tree, it is used as the root node of this decision tree, classifying over half of the negative examples based on whether the subject has a low severity of miscellaneous symptoms (nms_d9 < 7.5)[4]. This could be an indication that nonmotor-predominant PD patients do indeed have a wide manifestation and variety of nonmotor symptoms. After classifying on nms_d9, the tree then classifies negatiave examples as having rigidity ≥ 3.5, an example of how subtype 2 patients have relatively low motor symptoms. Finally, the tree classifies on the nonmotor symptom with the most information gain, nms_d2, where patients ≥ 7.5 are classified as falling into subtype 2.

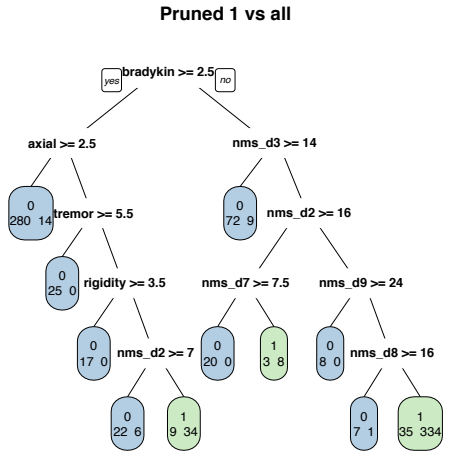


Figure 9: Cluster 1 (mild) vs all

5.1.3 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 several more nodes verifying the severity of rigidty, tremor, and axial, and the mildness of nms_d7 (urinary).

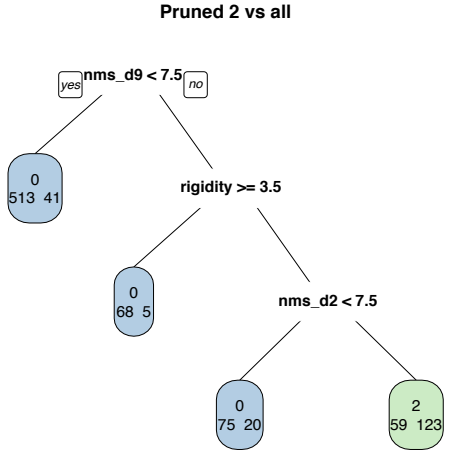


Figure 10: Cluster 2 (nonmotor-dominated) vs all

5.1.4 4 (severe)

The OVA tree for patients severely affected in all areas is predictable, testing entirely on whether or not symptoms (both motor and nonmotor) are relatively severe. Positive 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 2 shows, nms_d4 is perhaps the most distinguishing symptom of subtype 4 against nonmotor-predominant subtype 2 in particular, as subtype 2 has relatively mild percep/hallucination symptoms, in contrast to the comparable levels of severity for other nonmotor symptoms in both groups. 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.

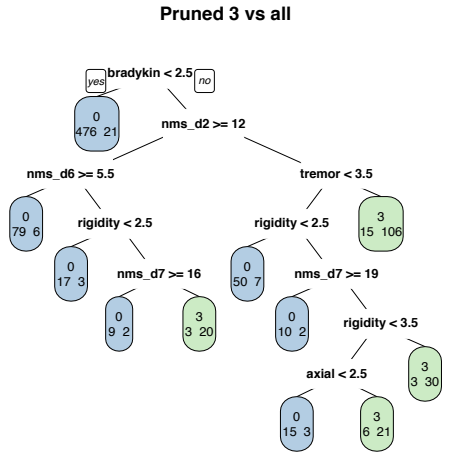


Figure 11: Cluster 3 (motor-dominated) vs all

[4]Recall that $0 <$ nms_d9 < 48.

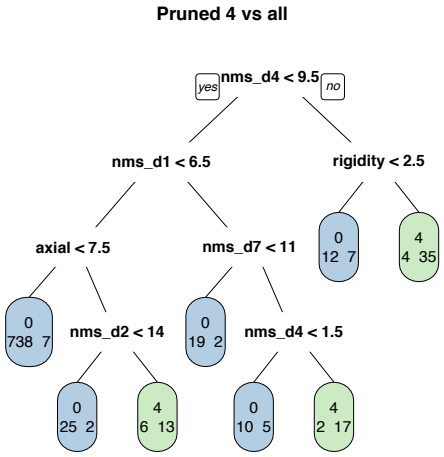


Figure 12: Cluster 4 (severe) vs all

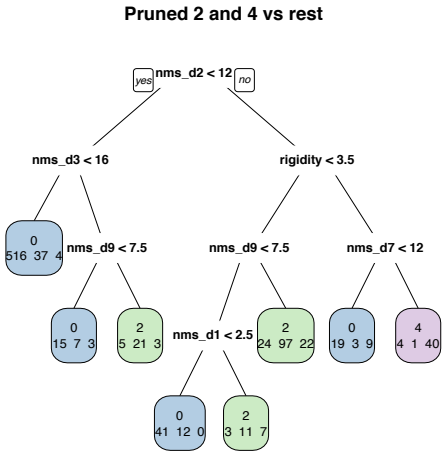


Figure 13: Clusters 2 (nms) and 4 (severe) vs rest (1 and 3)

5.2 Different angles of exploration: 2 and 4, 2 and 3 vs rest

There are many more interesting questions to be asked when examining the relationship between these clusters. One thing that may be helpful in understanding the relationship between the clusters is exploring different groupings of clusters for decision trees. The trees in Figure 13 and 14 are preliminary examples of this kind of exploration.

5.2.1 2 and 4 vs rest

In this tree, the node classifying examples as subtype 4 is localized to the furthest right branch. Predictably, examples in this node have scored relatively higher in rigidity ($\geq$ 3.5). Interestingly, a classification decision that is replicated in the 4 versus all decision tree is the decision to use nms_d7 (urinary) as a node, where subtype 4 is classified as having relatively high nms_d7 components ($\geq$ 12). Indeed, as shown in Figure 2, the mean of nms_d7 severity is similar to nms_d4 in that it is especially higher in cluster 4 than in cluster 2.

5.2.2 2 and 3 vs rest

In this tree, classification of patients in subtype 3 is primarily dependent on asserting bradykin $\geq$ 2.5 then splitting on tremor $<$ 3.5 is considered. Interestingly, the two nonmotor symptoms that differentiate cluster 3 are nms_d5 (attention/memory) and nms_d7 (urinary). Additionally, nms_d2, a quite important symptom, does not appear in the classification tree for this task.

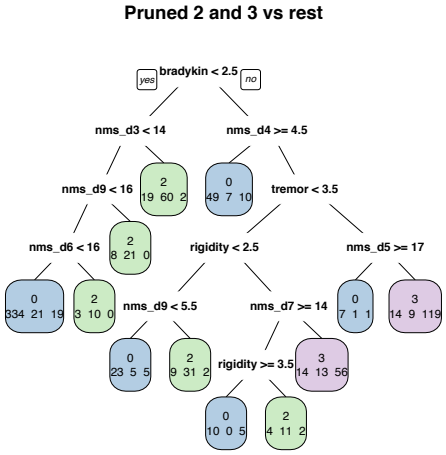


Figure 14: Clusters 2 (nms) and 3 (motor) vs rest (1 and 4)

5.3 Bayesian Networks

5.3.1 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.

5.3.2 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.

6 Longitudinal Analysis

I wanted to observe the progression of nonmotor symptoms of Parkinson's disease akin to work done by various studies (e.g. [2, 3, 4]). The data was a little bit too noisy when I tried to examine correlations, so I decided to average the values into bins for `durat_pd`: 0-2 years, 2-4 years, 4-6 years, etc. First, the correlation based on PD duration for each symptom is displayed in Figure 15.

> Note: I have not included all of the graphs I have, just those I found interesting. If you would like all of the graphs, I can send those to you in a separate email.

Most symptoms have at least a minor positive correlation. Interestingly, tremor, nms9 (anxiety) and nms10 (depression) do not have a strong correlation with PD duration.

6.1 Anxiety

I plotted the mean nms_9 (anxiety) score for each durat_pd bin in Figures 16 and 17

6.2 Depression

I plotted the mean nms_10 (depression) score for each durat_pd bin in Figures 18 and 19.

6.3 Cisitot

Note that cisitot has the highest correlation of all of the PD symptoms. I plotted cisitot per subtype in Figure 20. Interestingly, subtypes 2, 3, and 4 generally start from the same mean motor symptom score, suggesting different paths of disease progression.

6.4 Tremor

I plotted tremor symptoms in Figure 21. Notice how in this group, subtype 3 (motor-dom) is higher than all groups towards the beginning of disease progression. This supports the idea that Subtype 3 is Ma's [6] tremor-dominant cluster.

7 Extended Nonmotor Symptoms

7.1 Adding nonmotor symptoms to original k-means clustering

First, I used the clustering with the original 9 nms symptoms, but plotted the original 30 symptoms instead. The reason for this was to treat the clustering of the 9 symptoms as a form of dimensionality reduction, since we know each symptom of nms{1-30} is colinear with several other symptoms, and they form the broader nms_d{1-9} symptoms. The results are in Figure 22.

7.2 *New* clustering with 30 symptoms

7.2.1 Principal Component Analysis

I ran PCA on the dataset with solely the 30 nonmotor symptoms. A plot of various metrics for determining the optimal number of eigenvectors is in Figure 23. Most metrics agree that the optimal number of eigenvalues is around 6, but notice how one component has by far the highest eigenvalue; this component can be interpreted as a general "nonmotor severity" measure, indicating that much of the general nonmotor symptomatic expression falls along the same dimension.

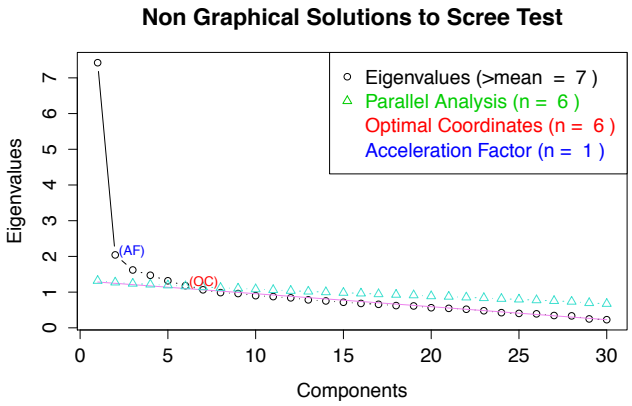


Figure 23: Eigenvalues and other metrics for Principal Component Analysis applied to the NMS30 data.

This principal component can be shown more clearly in the correlation plot of each symptom with the first 5 components located in Figure 24. In the first component, all of the symptoms have at least a minor positive correlation. The remaining components are composed primarily of dimensionality reduction in domain 3 (mood/cognition), domain 5 (attention/memory), domain 7 (urinary), and domain 9 (misc). It is once again demonstrated that mood/cognition symptoms are very important parts of this dataset.

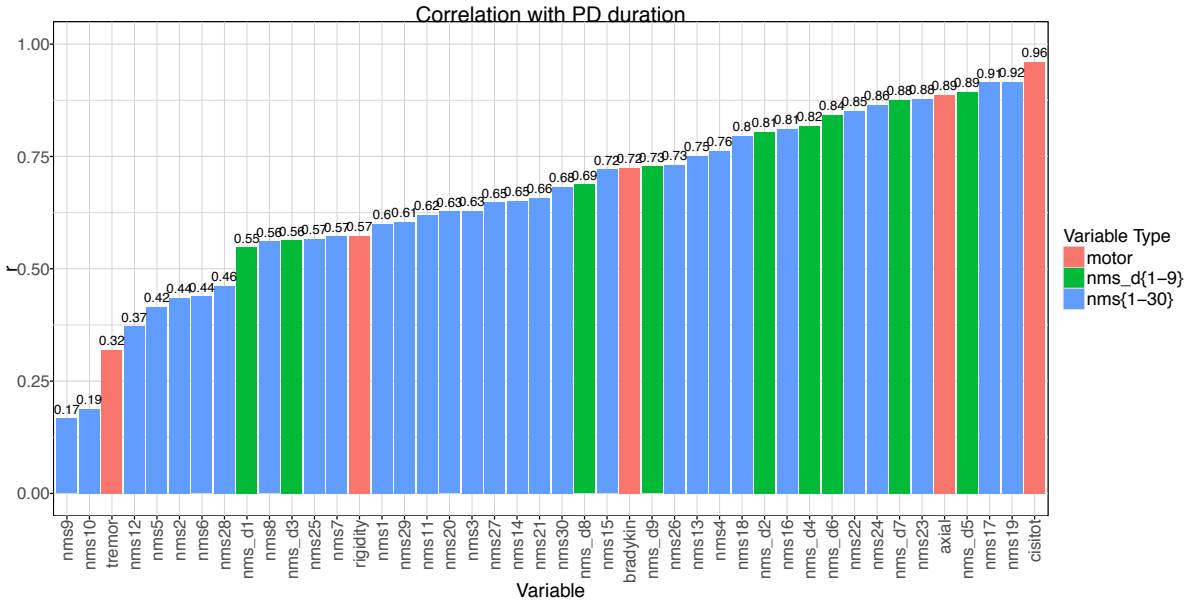


Figure 15: Correlation of variables according to PD Duration

7.2.2 Symptoms clustering

Nonmotor only I used hierarchical clustering to cluster the 30 symptoms of the dataset using complete linkage. Results are in Figure 25. Naturally, symptoms that belong to the same domain are clustered together. But the hierarchical clustering also shows how some symptomatic domains are closely related; for example, d6 (gastro), d7 (urinary), and d8(sexual). Additionally, miscellaneous symptoms are quite distinct from the other symptoms.

With motor symptoms In Figure 26, the results are similar to the original hierarchical clustering; but notice that motor symptoms are clustered together *except* for tremor. In fact, tremor is the most dissimilar symptom of them all!

7.2.3 k-means

More so than previously, metrics for identifying the number of clusters proved inconclusive. Most metrics suggested 2 clusters, which isn't a useful clustering. I decided to stick with $k = 4$ to compare the results of this new clustering with the previous run. All 30 NMS symptoms were included as well as the four motor symptoms axial, rigidity, bradykinesia, and tremor. In this clustering, 4 slightly different subtypes are identified:

1. ($n = 509$) Mildly affected in all domains.
2. ($n = 97$) Higher than average nonmotor symptoms, but severely affected in depressive symptoms especially.
3. ($n = 249$) Averagely affected in all domains.
4. ($n = 49$) Severely affected in all domains.

Because the number of symptoms is high, I don't provide the boxplots I have previously provided; instead, I provide a heatmap of z-scores in Figure 27 and a table of results in Table 8.

Using one-way ANOVA for multiple means, only one variable had no significant interaction with the cluster ($p > 0.05$): pdonset.

7.2.4 Model-based expectation-maximization

Because the dimensionality is quite high in this case; k-means may not have provided particularly accurate results; thus, I tried using a Gaussian mixture model-based clustering package. Various models were tried, with Bayesian Information Criterion (BIC) plotted in Figure 28. The best fitting model(s) that maximized the Bayesian Information Criterion include the diagonal, varying volume, equal shape Gaussian mixture model (VEI) at 11 clusters and the ellipsoidal, equal shape model (VEV) with 3 clusters. However, the VEV model wasn't interesting; it simply partitioned individuals into low, medium, and high overall severity.

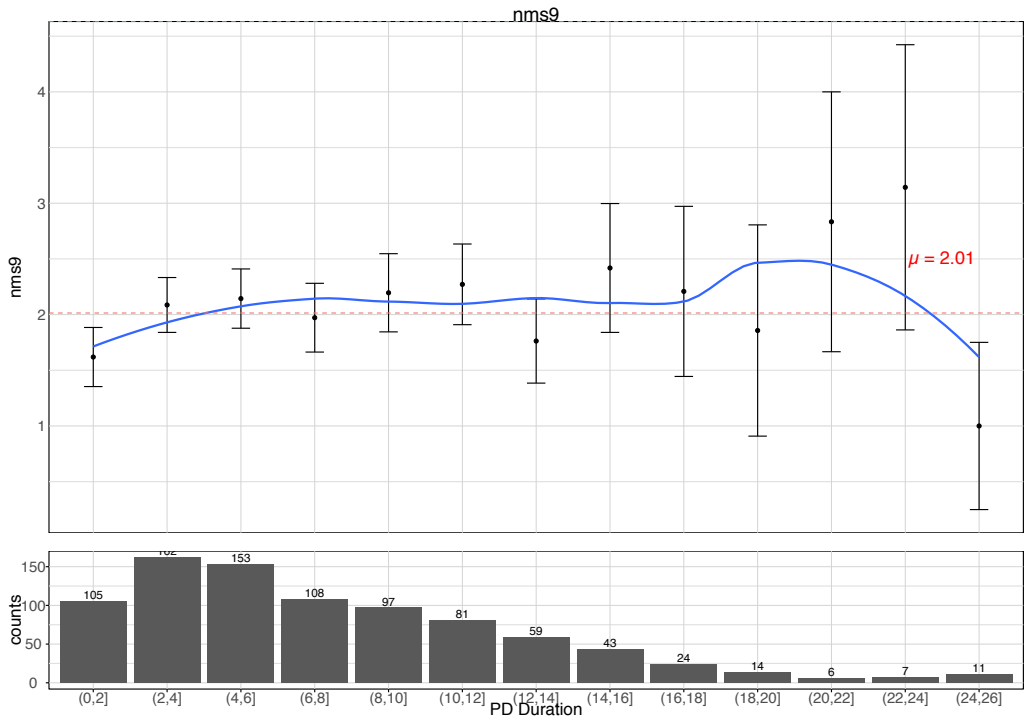


Figure 16: Mean anxiety (nms_9) score by PD duration, with number of patients in each group at bottom.

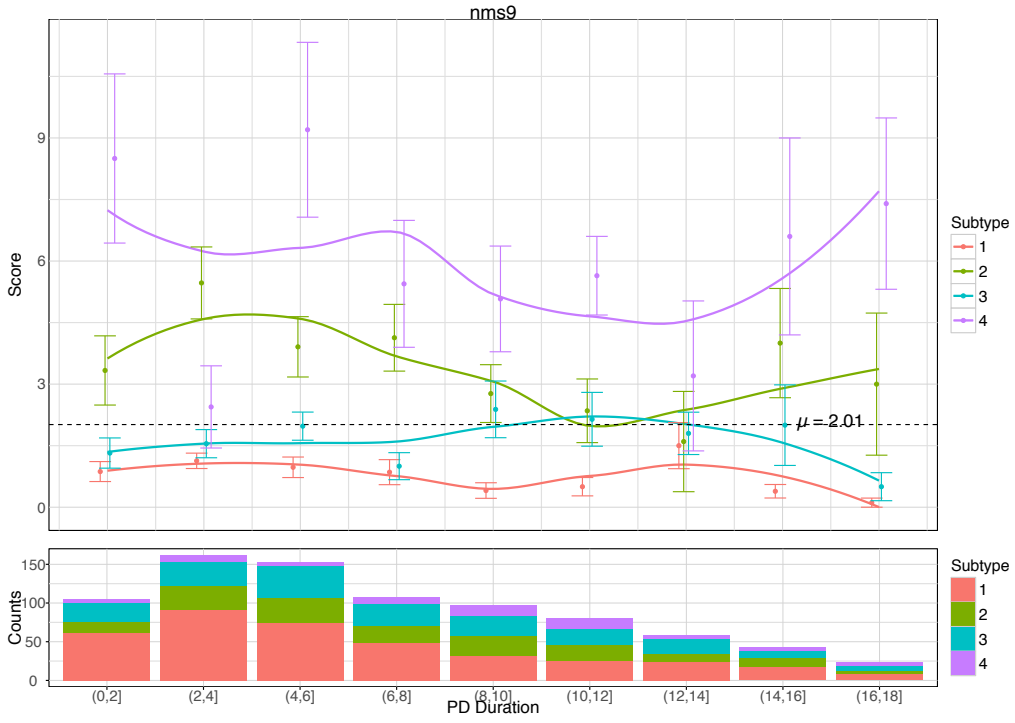


Figure 17: Mean anxiety (nms_9) score per subtype, with number of patients in each group at bottom.

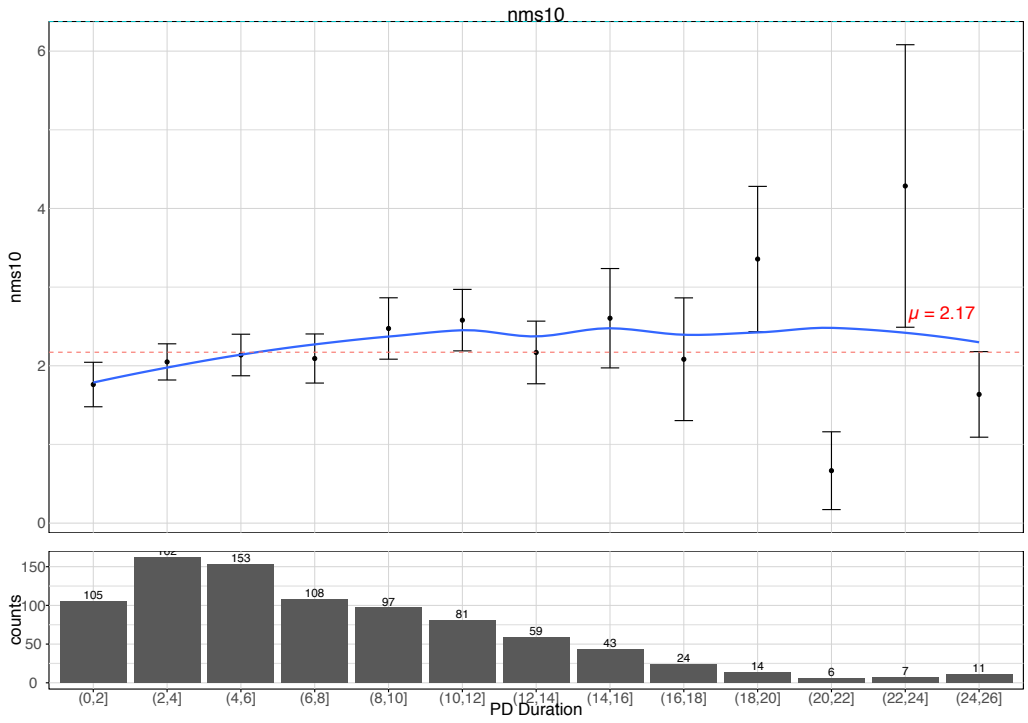


Figure 18: Mean depression (nms_10) score by PD duration, with number of patients in each group at bottom.

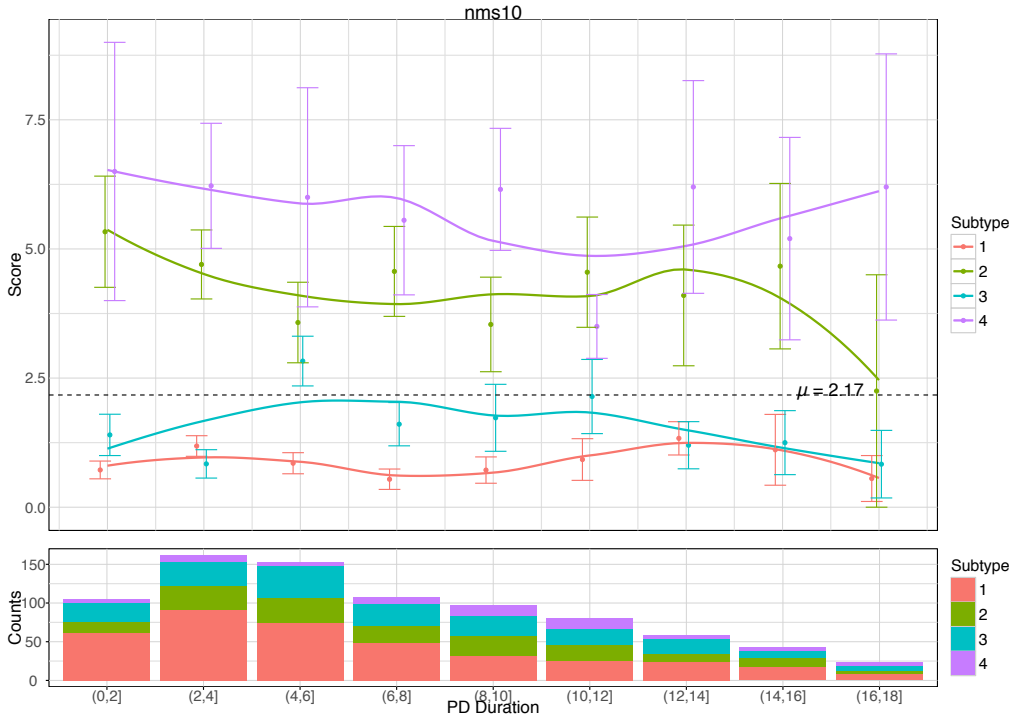


Figure 19: Mean depression (nms_10) score per subtype, with number of patients in each group at bottom.

Table 8: Mean symptom scores for the centers of the k-means clustering on nms{1-30} where $k = 4$. In this case, a Depression-Dominant subtype has been identified.

cluster	1	2	3	4
d1-1-lightheaded	2.08	2.43	0.55	5.18
d1-2-fainting	0.17	0.53	0.13	2.45
d2-3-drowsiness	2.51	2.77	0.97	7
d2-4-fatigue	4.84	6.29	1.3	7.31
d2-5-insomnia	3.16	5.46	1.2	4.22
d2-6-rls	1.88	2.31	0.66	3.88
d3-7-loss_interest	1.04	**6.08**	0.41	4.59
d3-8-loss_activities	1.7	**7.14**	0.7	5.37
d3-9-anxiety	1.97	**6.63**	0.91	4.57
d3-10-depression	2.35	**7.81**	0.81	4.29
d3-11-flat_affect	0.91	**4.95**	0.39	2.71
d3-12-loss_pleasure	1.09	**6.82**	0.36	3.43
d4-13-hallucination	0.58	0.82	0.2	4.63
d4-14-delusion	0.16	1.49	0.11	3.14
d4-15-diplopia	0.56	1.01	0.16	4.39
d5-16-loss_concentration	2.47	4.02	0.91	6.55
d5-17-forget_explicit	2.18	3.77	0.86	6.69
d5-18-forget_implicit	1.76	3.12	0.68	6.65
d6-19-drooling	3	2.99	0.67	6.1
d6-20-swallowing	1.74	1.24	0.37	4.53
d6-21-constipation	3.31	3.63	1.65	6.96
d7-22-urinary_urgency	3.84	3.49	0.99	7.8
d7-23-urinary_frequency	3.79	3.98	0.94	6.51
d7-24-nocturia	5.13	4.99	1.67	7.65
d8-25-sex_drive	2.34	4.46	0.76	5.16
d8-26-sex_dysfunction	2.39	3.21	0.79	4.61
d9-27-unexplained_pain	2.98	2.48	0.7	4.24
d9-28-gust_olfact	3.18	3.57	1.49	4.73
d9-29-weight_change	1.86	2.54	0.88	3.92
d9-30-sweating	2.77	2.2	0.68	3.18
tremor	2.59	2.12	2.53	4.1
bradykin	2.83	2.76	1.98	3.86
rigidity	2.61	2.42	1.87	3.61
axial	4.22	4.4	2.17	7.24

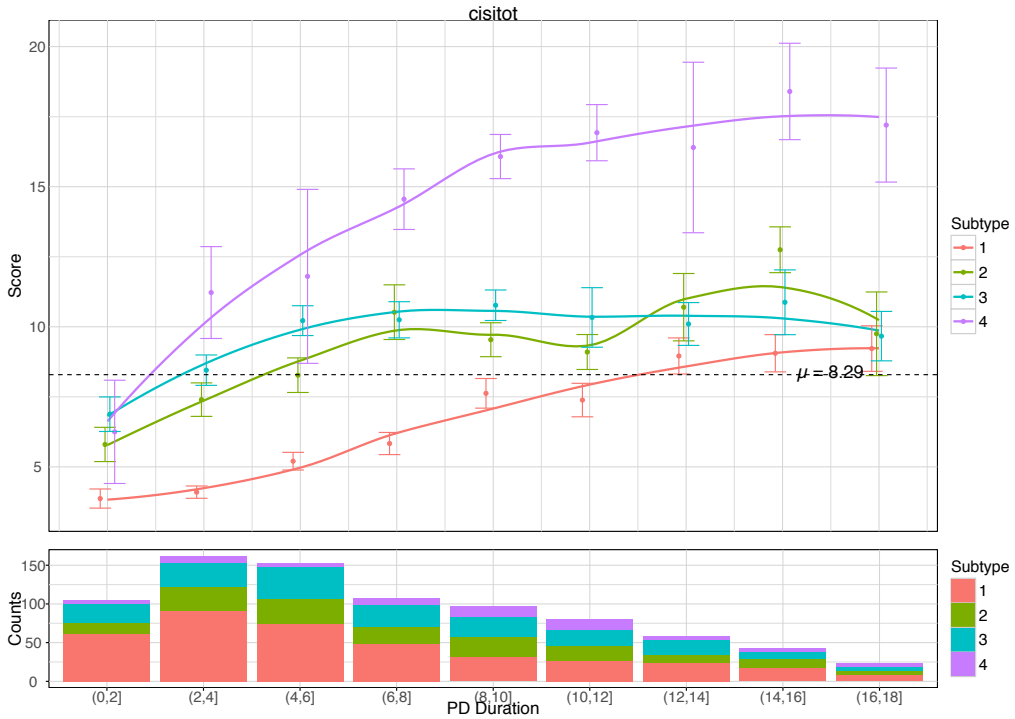


Figure 20: Mean cisitot score per subtype, with number of patients in each group at bottom.

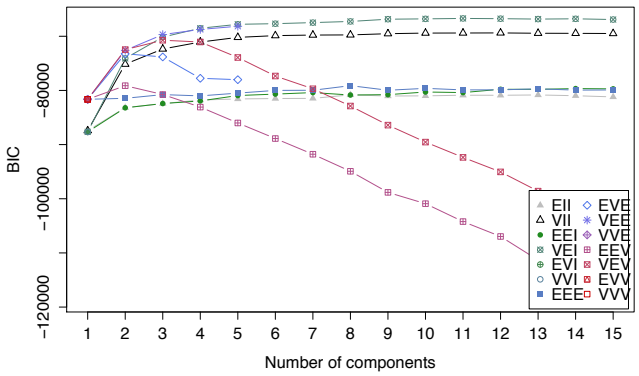


Figure 28: Bayesian Information Criterion for various Gaussian mixture models, plotted against the number of clusters in each model. The best model found is the diagonal, varying volume, equal shape Gaussian mixture model (VEI) at 11 clusters; notice also that the ellipsoidal, equal shape model (VEV) is able to achieve a relatively high BIC with only 3 clusters. The VEV model, however, is not particularly informative.

The VEI model, with 11 clusters, is more interesting. There are many clusters in the data that are expected, including a few generally mild clusters and a few generally severe clusters. But, by isolating more "interesting" clusters, we begin to see smaller, more specific proportions of PD patients. These clusters seem to clarify and expand upon the results found in previous analyses. The clusters, summarized, are:

1-3. ($n = 41,\ 77,\ 150$) Mild in all domains. Cluster 1 is especially mild in motor symptoms (see cisitot).

4. ($n = 38$) Insomnia-dominant.

5. ($n = 43$) Motor-dominant.

6. ($n = 163$) Average.

7. ($n = 54$) Urinary-dominant.

8. ($n = 177$) Above-average in all domains.

9. ($n = 67$) Depression-dominant.

10. ($n = 81$) Nonmotor-dominant.

11. ($n = 13$) Severe in all domains.

A heatmap of all 11 clusters found is located in Figure 29.

7.3 Decision Trees

I created decision trees for the k-means $k = 4$ clustering, located in Figure 30.

One vs all decision trees are located in Figures 31, 32, 33, and 34.

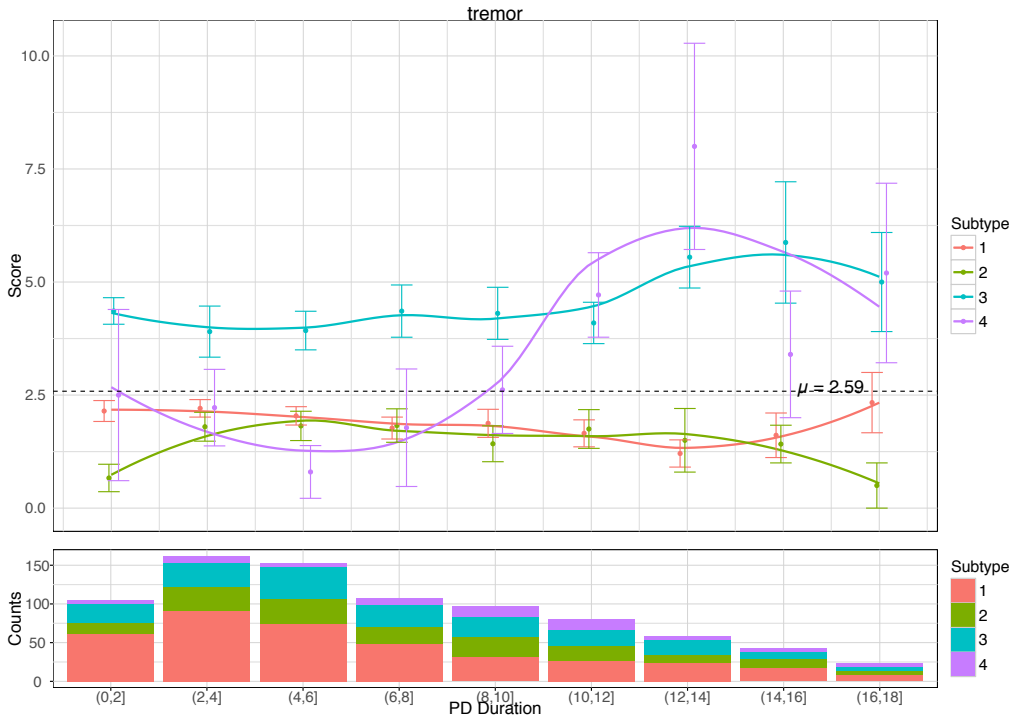


Figure 21: Mean tremor score per subtype, with number of patients in each group at bottom.

8 Preliminary Conclusions

8.1 Overall clustering

k-means clustering on this Parkinsons' Disease data set reveals clusters that confirm previous findings in the field, mainly van Rooden et al. [5] and the identification of four subtypes of Parkinson's disease: mild, nonmotor-predominant, motor-predominant, and severe. van Rooden's work was done with a separate dataset using a different modeling method (expectation-maximization), and this investigation independently confirms these subtype classifications. Unlike van Rooden, mean disease durations differences do exist between subtypes 1 (mild) and 4 (severe), likely due to further development of the disease, although the differences between 2 and 3 (nonmotor/motor predominated) subtypes are insignificant (Table 4), suggesting different developmental paths of the disease.

Overall, little information was found in pdonset, durat_pd, or current age, according to Tables 4 and 5. Mean ages were similar for subgroups 1, 2, and 3 ($p > 0.05$, but different for the severe subtype 4, which makes sense given that patients in 4 also have longer disease durations. Specifically, clusters 1 and 4 seem to be phenotypically quite similar, except at different stages of disease progression, given cluster 4's higher age and durat_pd scores.

However, clusters 2 and 3 clearly show different disease progression, one in the motor direction, and one in the nonmotor. Both groups have similar age, pdonset, and durat_pd scores, but differ wildly in symptomatic expression. Cluster 2 is dominated by a high prevalence of nonmotor symptoms, such as nms_d2, nms_d3, nms_d7, and nms_d9. Cluster 3, however, is dominated by a high prevalence of motor symptoms, while most motor symptoms are similar to the mild cluster 1. Of note is that the tremor population mean is the highest cluster mean, even higher than the severe subtype 4. This motor-dominant cluster may thus overlap with Ma's tremor dominant/slow progression cluster [6].

Generally, given stable pdonset scores and predictably increasing durat_pd scores for clusters 1 and 4, Ma et al's rapid disease progression/late onset and tremor dominant/slow progression clusters [6] were mostly not found in this dataset, save for the tremor-dominant motor cluster.

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] and are similar to Sauerbier's identification of sleep dominant and cognitive dominant clinical NMS subtypes [7]. Compared to Erro et al. [8], nonmotor/motor dominant subtypes were indeed found, but an additional subgroup with relatively severe levels of both motor and nonmotor symptoms were found. Erro's benign subtype groups possibly overlap with the mild cluster 1 found in this investigation.

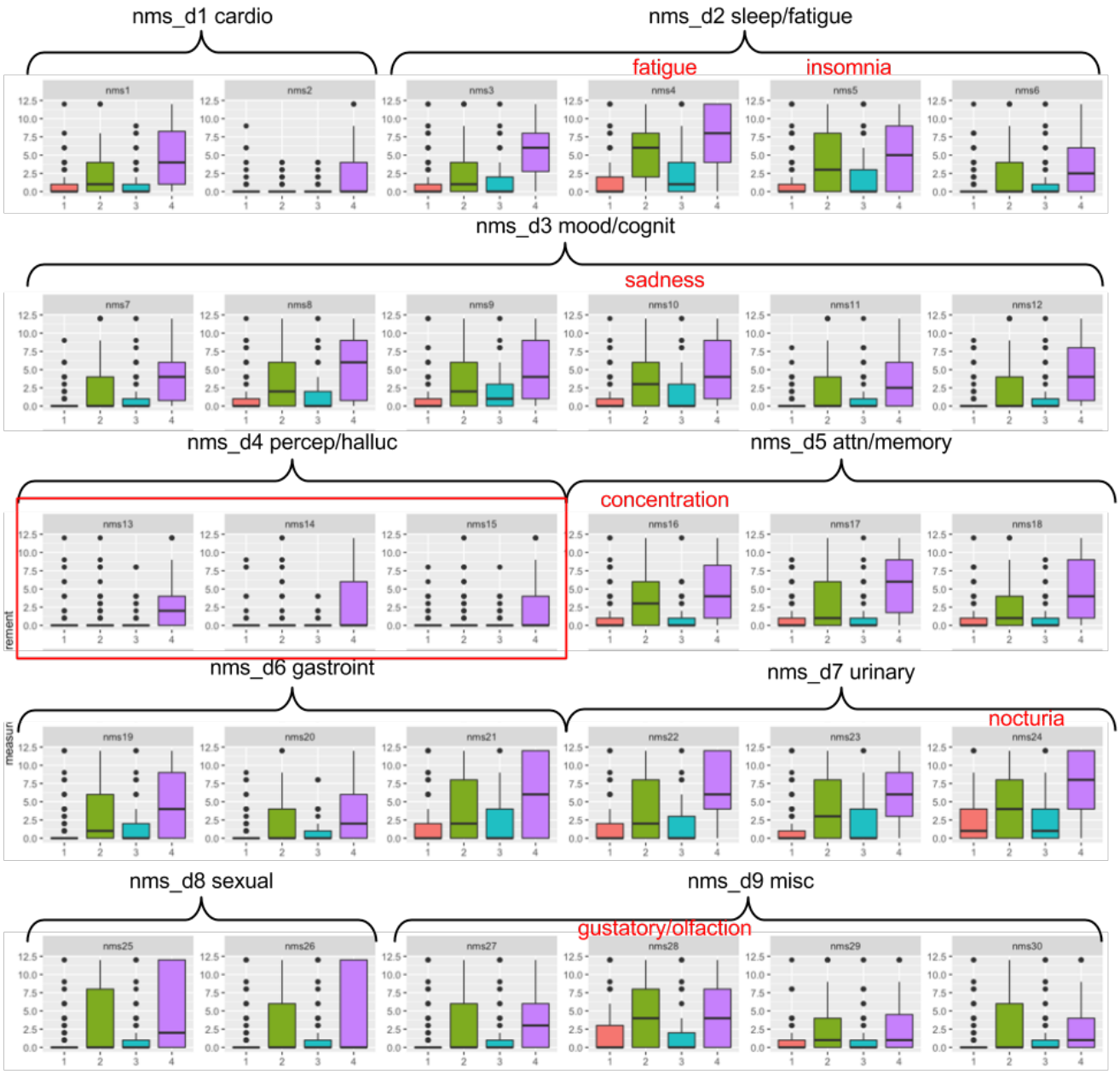


Figure 22: All symptoms (nms{1-30}) plotted with the original clustering results of the the 9 broader symptoms. Interesting parts of the figure are are labeled in red.

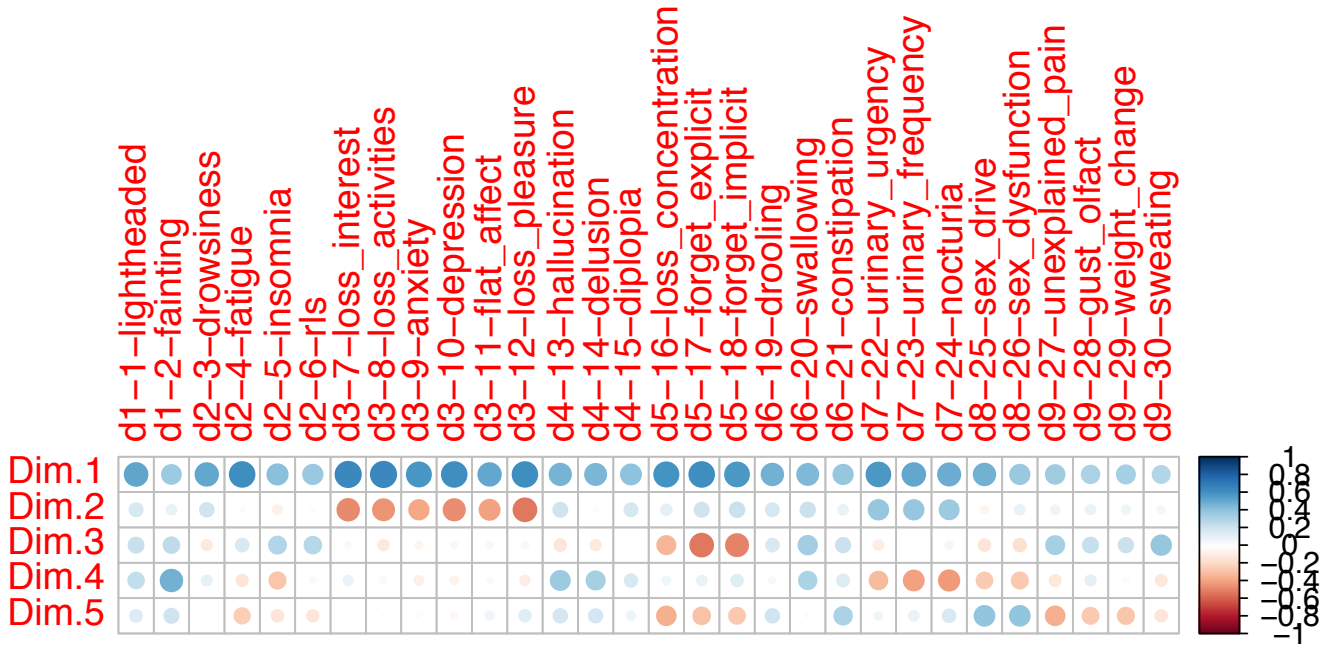


Figure 24: Correlation of each nonmotor symptom with the first 6 components

8.2 Nonmotor subtype: clustering and modeling

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. In k-means subdivision of the nonmotor-dominant subtype where $k = 2$ and $k = 3$, opposite trends were confirmed with nms_d2 and nms_d3 symptoms. Similarly, in the 2 and 4 vs rest decision tree (Figure 13), nms_d2 and nms_d3 nodes were used to differentiate various categories of nomotor-dominant patients. When $k = 3$, the subtype with the highest nms_d2 scores and lowest nms_d3 scores had by far the highest axial scores, nms_d6 (gastrointestinal) scores, and nms_d7 (urinary) scores. Thus subtype 3 of the nonmotor-dominated group could include patients falling into the cognitive/depression-dominant or autonomic dominant subtypes.

Despite the variety in symptomatic expression in this nonmotor group, what seems most consistent is the presence of nms_d9 (miscellaneous) nonmotor symptoms, as it is used as the root node of the 2 vs all decision tree (Figure 10) and the 2 and 4 vs rest decision tree (Figure 13).

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

8.3 New conclusions

First, the longitudinal analysis gives more insight into the clusters found when clustering on nms_d{1-9}. According to Figure 15, most symptoms are highly correlated, with PD duration, but notably, mood/cognition symptoms (nms9, nms10, nms12) and tremor are not correlated highly with PD duration. The differences in disease progression can be seen by the corresponding graphs, Figures 17 and 19. In both graphs, what is interesting is that Subtype 2 (Nonmotor-Dominant) starts at higher scores for nms_9 and nms_10, thus indicating that these patients' subtype is can be determined early in PD duration from depressive symptom score. Similarly, when examining Subtype 3 (Motor-dominant) in Figure 21, the mean tremor score is substantially higher from PD onset. Interestingly, Subtype 4 (Severe) generally starts at lower tremor and motor scores during disease onset (Figure 20), but then rises sharply, exceeding other Subtypes. More evidence that tremor is a unique motor symptom is located in Figure 26, where it is the most distant symptom from all other symptoms.

When examining all 30 symptoms, more evidence is given that the previously-discovered Subtype 2 (Nonmotor-Dominant) may be primarily characterized by

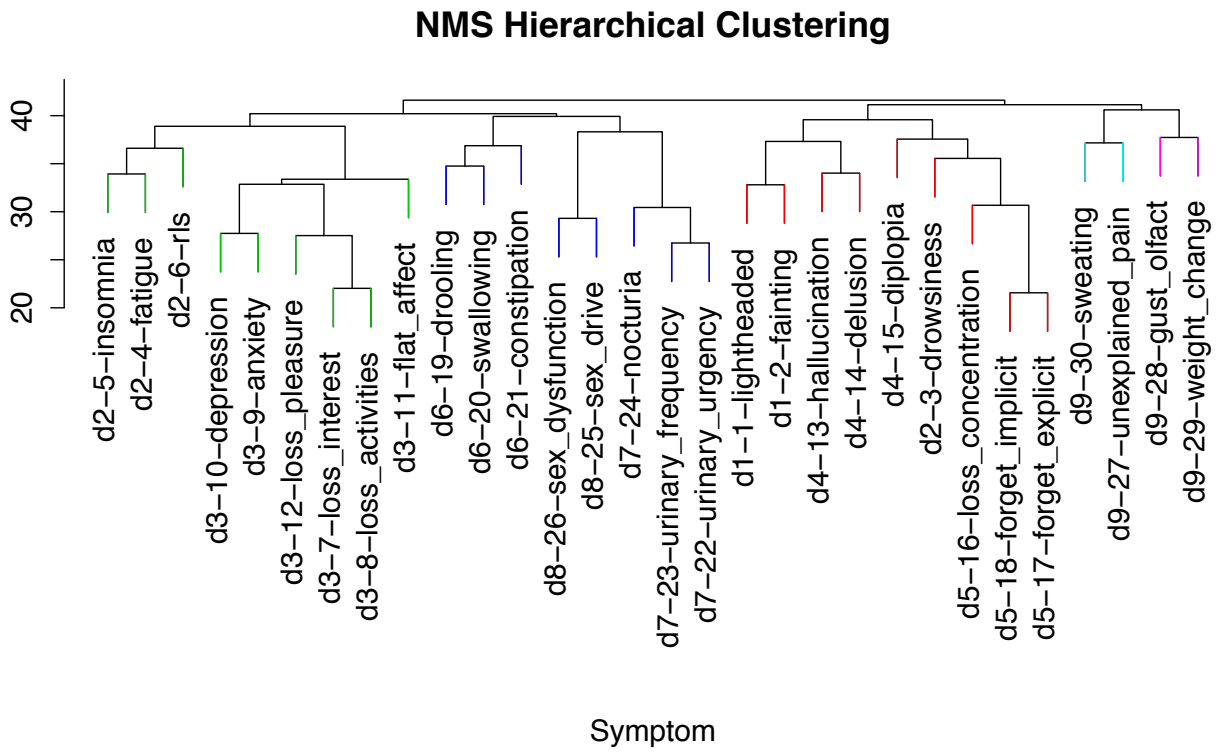


Figure 25: Hierarchical clustering of symptoms. Dendrogram colored with 5 clusters. Naming is: d{1-9}-{1-30}-description

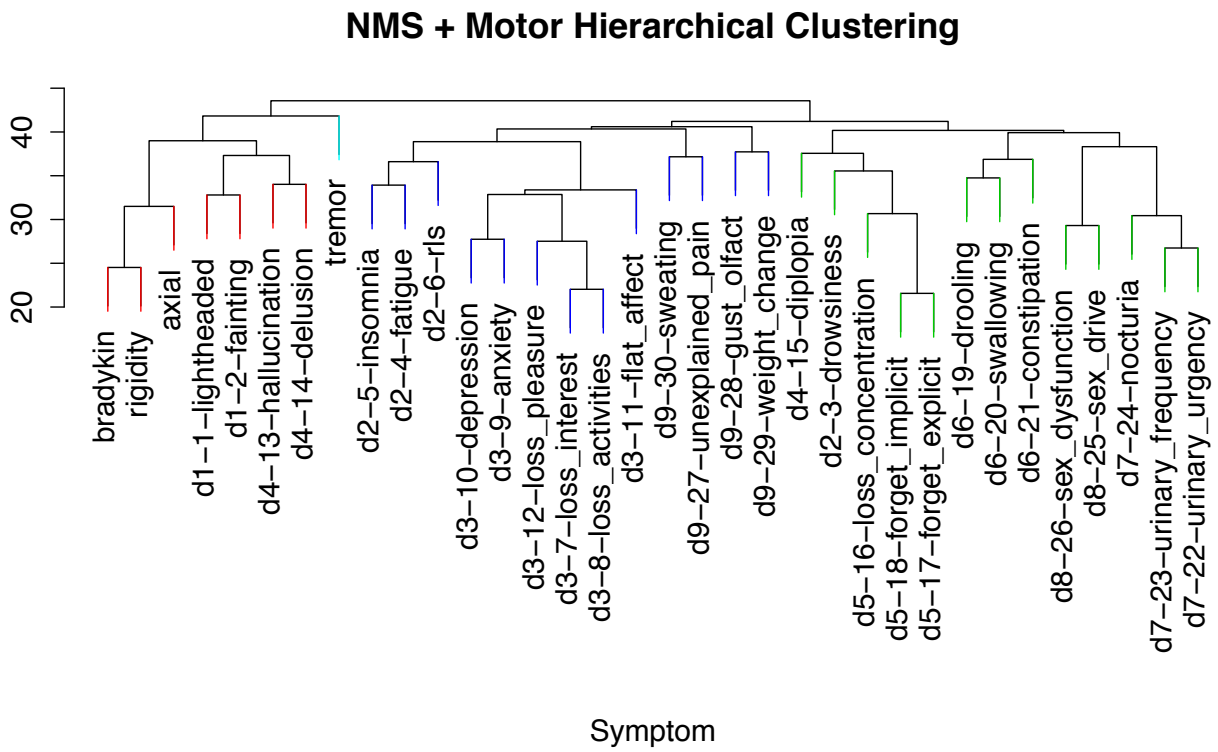


Figure 26: Hierarchical clustering of symptoms including motor symptoms. Dendrogram colored with 4 clusters.

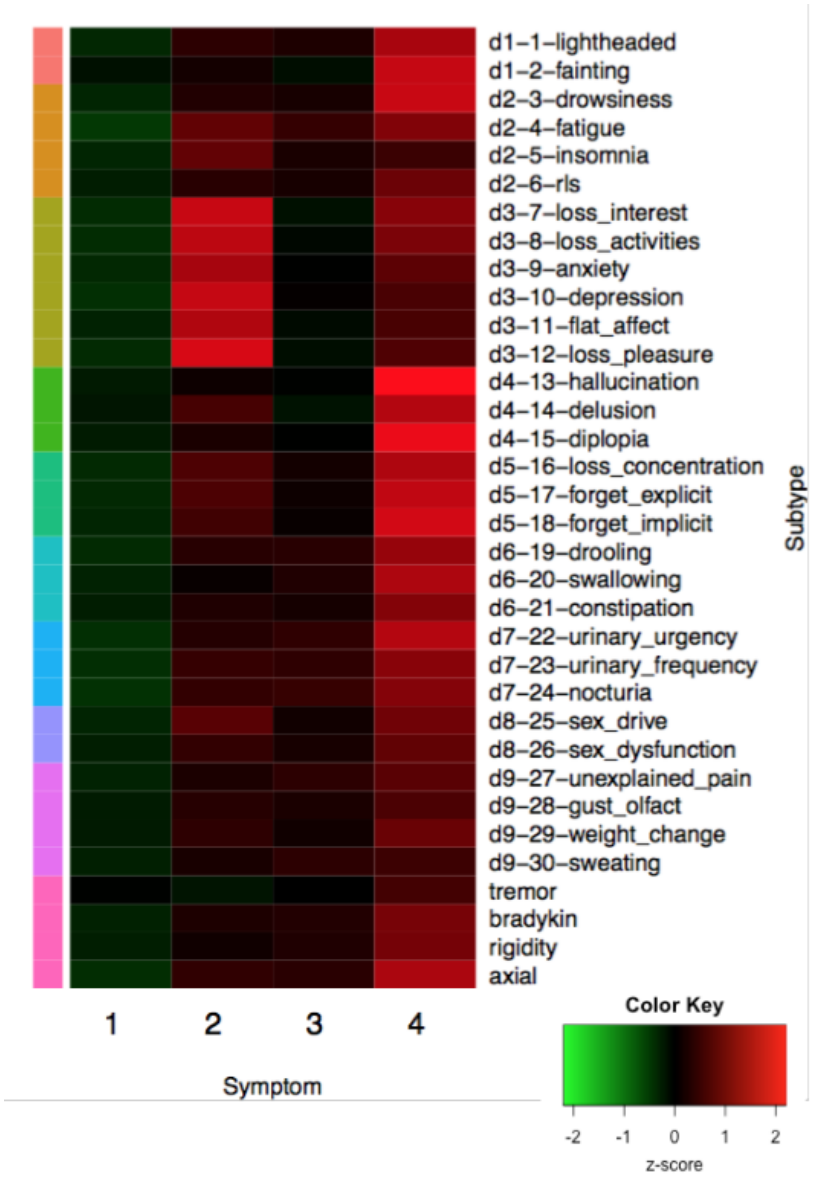


Figure 27: Heatmap for k-means clustering, $k = 4$, with the 30 nonmotor symptoms. Notice that subtype 2 is a "Depression-Dominant" subtype.

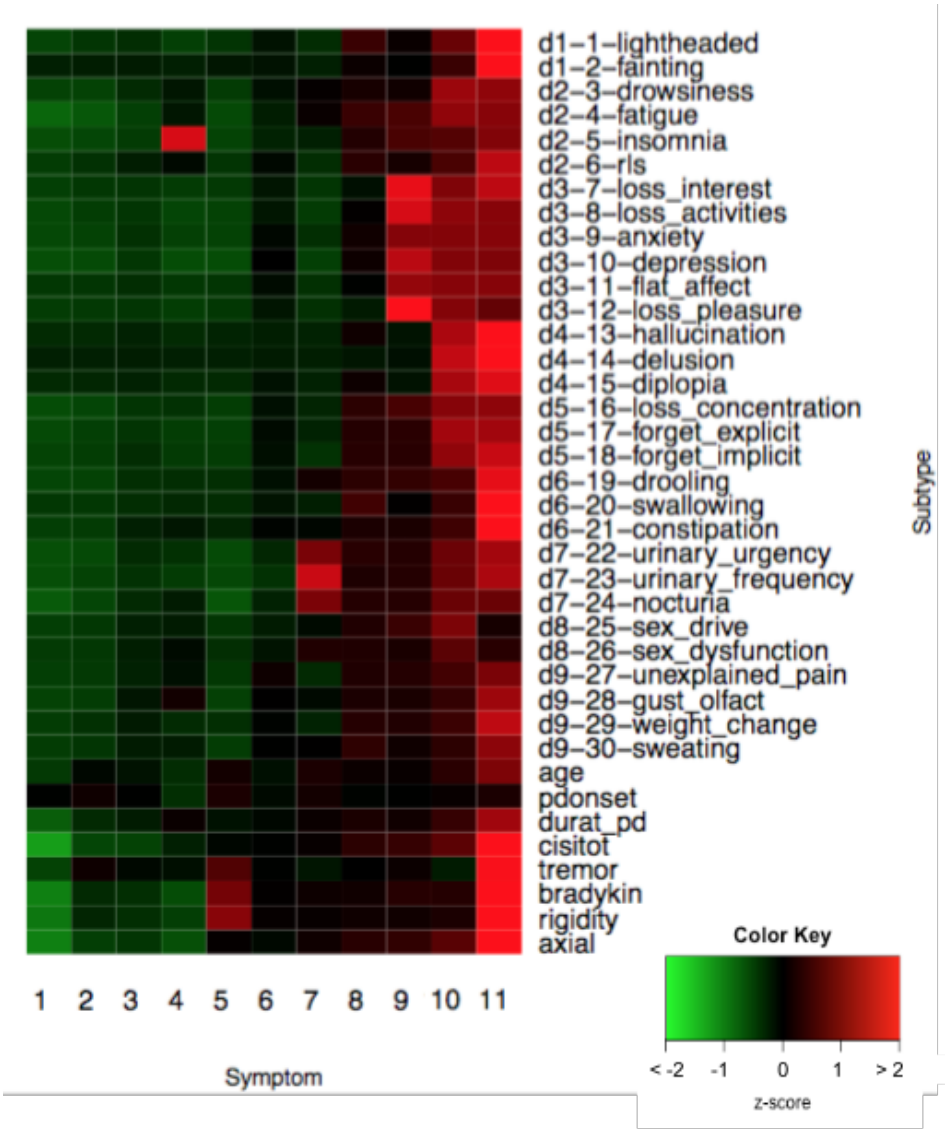


Figure 29: Heatmap of clusters found using the VEI Gaussian Mixture Model.

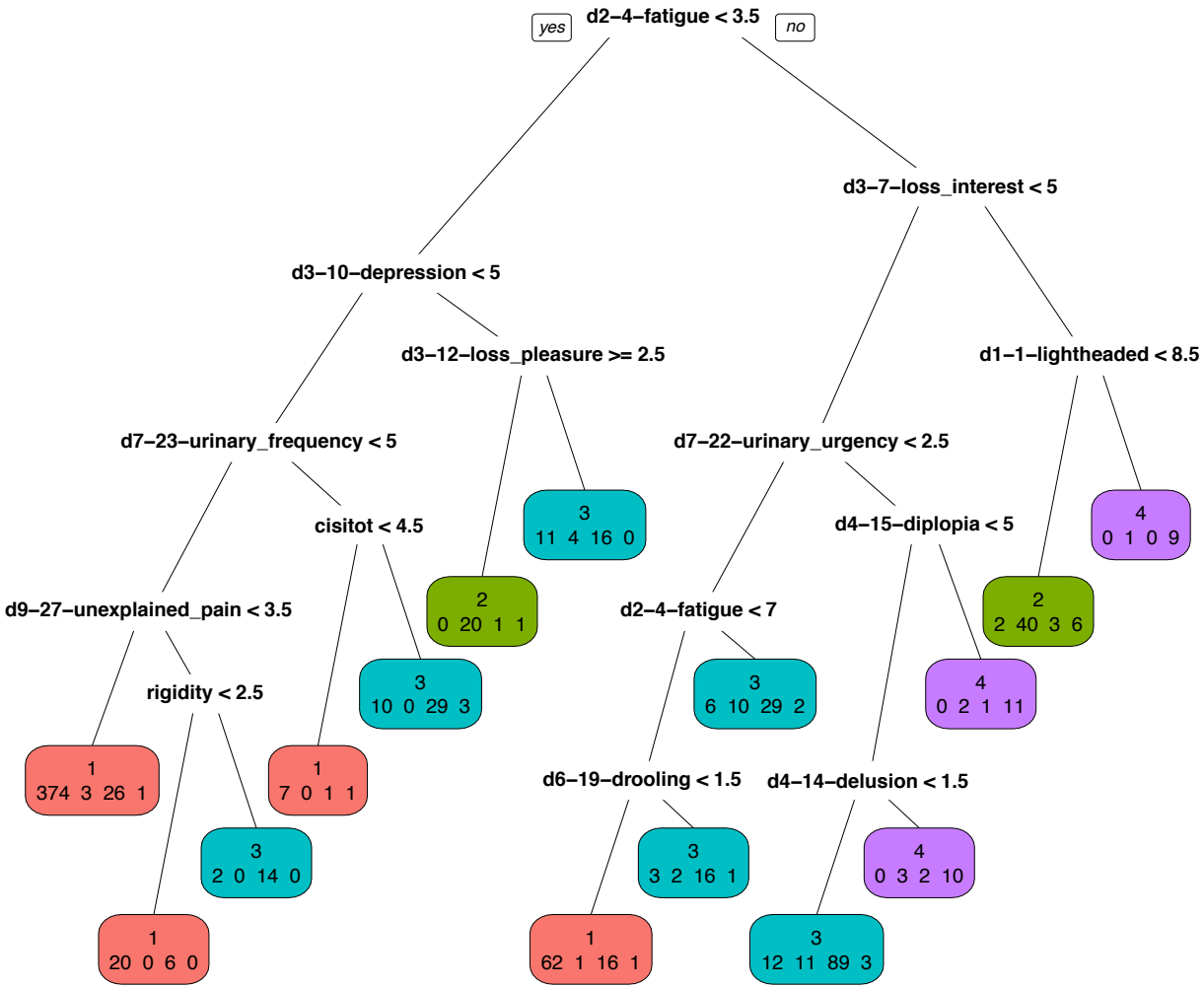


Figure 30: Pruned decision tree, clustered on nms{1-30}

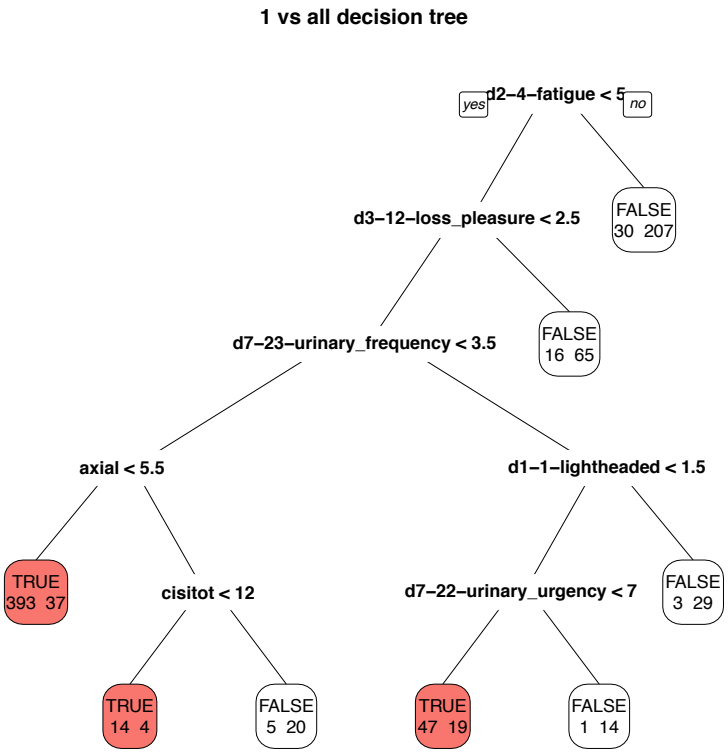


Figure 31: Pruned 1 vs all decision tree.

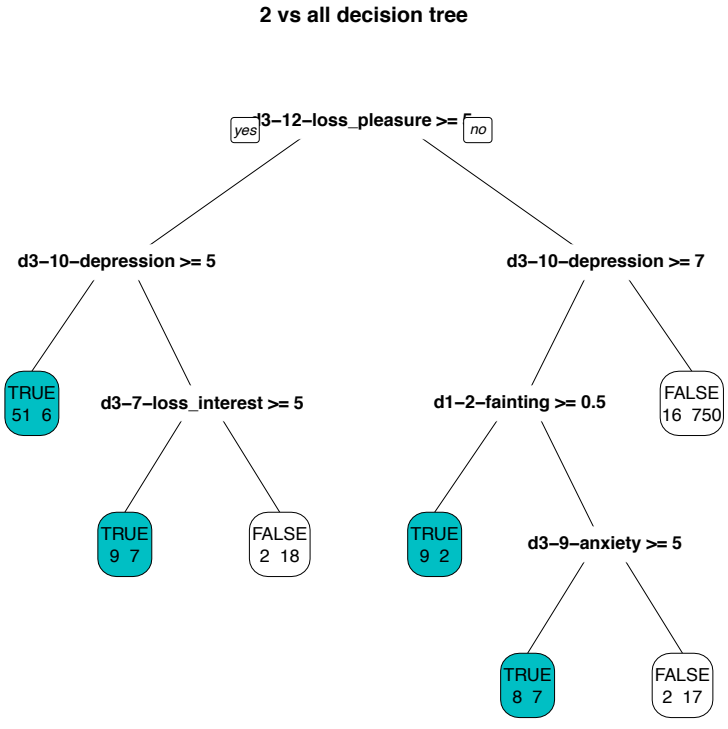


Figure 32: Unpruned 2 vs all decision tree. (Pruned was too simple).

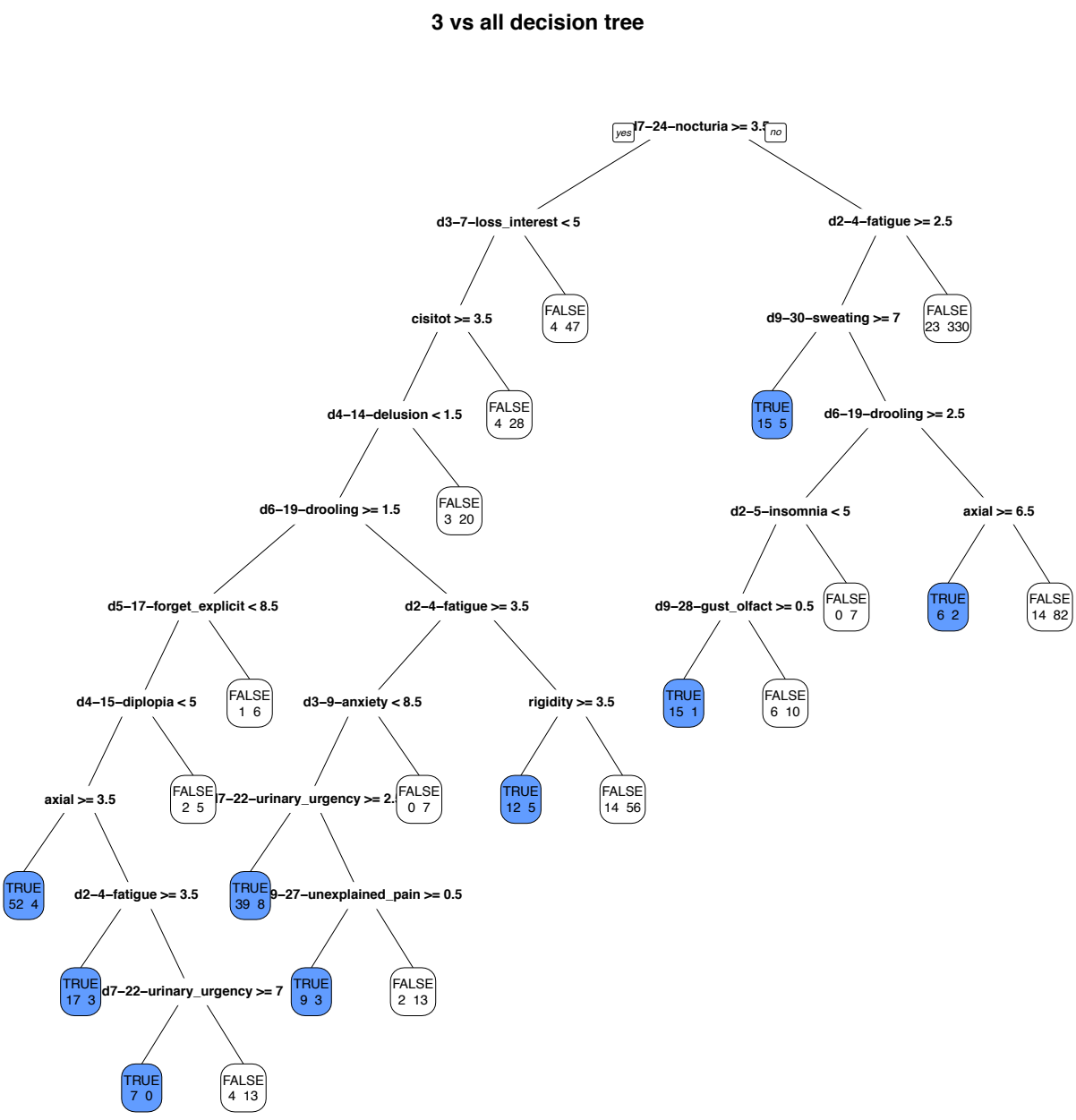


Figure 33: Pruned 3 vs all decision tree.

high depressive symptoms. By graphing the 30 subtypes attached to the original clustering, as in Figure 22, the mean scores of nms8, nms9, and nms10 for Subtype 2 are substantially higher than Subtypes 1 and 3.

Indeed, PCA on the 30 nonmotor symptoms identifies the second-most prominent component as a general mood/cognition component, and k-means clustering on the 30 symptoms only (Figure 27) divides the 1000 patients into four slightly different groups, a mild, average, depression-dominant, and severe group.

The Gaussian mixture model identified in Figure 29 fragments the previously-discovered clusters into more groups. Here, a wide variety of specialized subtypes of PD are displayed, including insomnia, urinary, motor, nonmotor, and depression-dominant groups, as well as the expected mild, average, and severe subtypes. It is likely that the previous analysis with only nms_d{1-9} combined most of those specialized groups into the more general nonmotor-dominant Subtype 2.

It's intuitive that a Depression-Dominant group emerges when clustering on nms{1-30}, since domain 3 consists of 5 separate measures. Thus, any high expression of depressive symptoms is magnified in clustering, since the symptoms are highly similar (Figure 26) and treated with equal weight. Once again reinforcing what was discovered previously, depressive symptoms have been shown to be very important in determining subtypes of PD.

References

[1] S.-M. Fereshtehnejad, *et al.*, *JAMA neurology* **72**, 863 (2015).

[2] C. Pont-Sunyer, *et al.*, *Movement Disorders* **30**, 229 (2015).

[3] T. C. Vu, J. G. Nutt, N. H. Holford, *British journal of clinical pharmacology* **74**, 267 (2012).

[4] L. B. Zahodne, *et al.*, *Neuropsychology* **26**, 71 (2012).

[5] S. M. van Rooden, *et al.*, *Movement Disorders* **25**, 969 (2010).

[6] L.-Y. Ma, P. Chan, Z.-Q. Gu, F.-F. Li, T. Feng, *Journal of the neurological sciences* **351**, 41 (2015).

[7] A. Sauerbier, P. Jenner, A. Todorova, K. R. Chaudhuri, *Parkinsonism & related disorders* **22**, S41 (2016).

[8] R. Erro, *et al.*, *PLoS One* **8**, e70244 (2013).

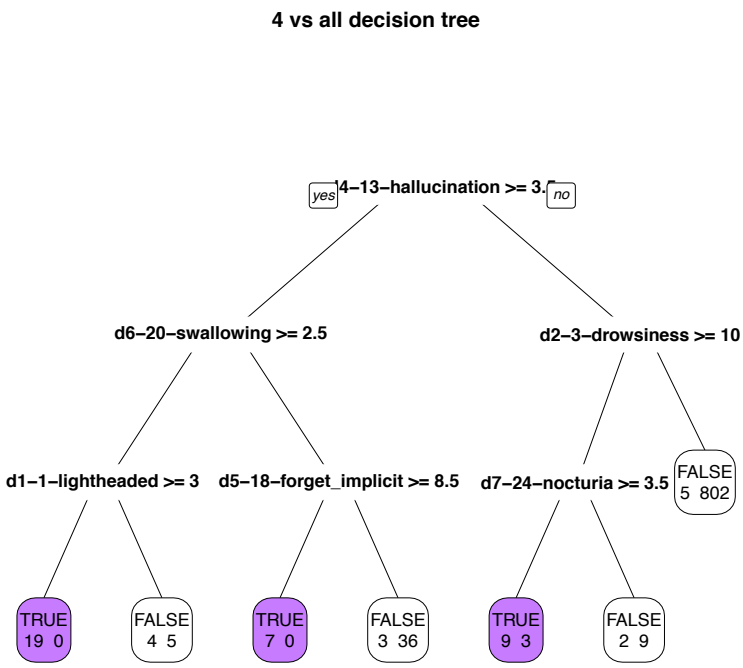


Figure 34: Unpruned 4 vs all decision tree. (Pruned was too simple).