Cluster Analysis: Identifying Parkinson's Disease Subtypes

July 16, 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). 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

2 k-means

k-means clustering with $k = 4$ was tried. $k = 2, 3$ provided models that were too simplistic. $k = 5$ did not provide any new information, but rather just fragmented existing groups.

Table 3: Cluster statistics

Cluster	n
1	79
2	394
3	275
4	153

2.1 Decision tree

k	CP[3]	CV Xerror[4]	Root Feature	Root Error	Figure
4	0.0100	0.255	pigd < 2.5	0.563	Figure 1

Table 4: k-kmeans decision trees statistics

2.2 Interpretation of Clusters

2.2.1 Cluster summaries

Available in Figure 2. Error bar is standard error.

[3]Complexity Parameter

[4]10-fold cross validation

2.2.2 Interpretation

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:

```
age insignificant differences:
         diff         lwr        upr        p adj
4-3 -2.271990  -4.7160927  0.1721117 7.920343e-02
sex insignificant differences:
           diff        lwr        upr     p adj
2-1 -0.09275204 -0.2454183 0.05991417 0.4000134
3-1 -0.14660529 -0.3046921 0.01148156 0.0803067
4-1  0.04757177 -0.1240051 0.21914864 0.8917054
3-2 -0.05385325 -0.1511666 0.04346014 0.4843788
pdonset insignificant differences:
           diff       lwr      upr     p adj
2-1 -0.90162565 -4.302019 2.498768 0.9037828
3-1 -0.05040276 -3.571532 3.470727 0.9999820
4-1 -0.53727145 -4.358869 3.284326 0.9837825
3-2  0.85122289 -1.316276 3.018722 0.7431486
4-2  0.36435420 -2.263251 2.991959 0.9844187
4-3 -0.48686869 -3.268952 2.295214 0.9695444
nms_d1 insignificant differences:
          diff         lwr       upr        p adj
3-2  0.5565021 -0.01586385  1.128868 6.018674e-02
nms_d3 insignificant differences:
          diff        lwr        upr       p adj
4-1  -1.192024  -4.493510   2.109461 0.789133802
nms_d4 insignificant differences:
          diff        lwr       upr        p adj
3-2  0.2951915 -0.3318339  0.922217 6.195406e-01
nms_d5 insignificant differences:
           diff         lwr       upr        p adj
3-2   0.7841071  -0.4814825  2.049697 0.3821565990
nms_d8 insignificant differences:
          diff       lwr      upr     p adj
4-1 -0.9010507 -2.846119 1.044017 0.6318563
3-2  0.7255838 -0.377602 1.828770 0.3280035
nms_d9 insignificant differences:
          diff         lwr       upr       p adj
4-1  0.7048896  -1.6510530  3.060832 0.867951568
tremor insignificant differences:
          diff       lwr        upr      p adj
```

```
3-1 -0.2662831 -1.064520 0.53195417 0.82613187
4-2 -0.5667198 -1.162396 0.02895624 0.06895758
rigidity insignificant differences:
          diff         lwr        upr        p adj
4-2  0.2310142 -0.03536748  0.4973959 1.153952e-01
```

2.2.4 Ranked Features by Information Gain

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

2.2.5 Correlation Plots

Figure 3.

2.2.6 Bradykinesia and rigidity

Figures 4 and 7

3 Other Work

3.1 Bayesian Networks

In Figure 8. Structure is too sparse, need to discretize or use some kind of regularization (e.g. a lasso)

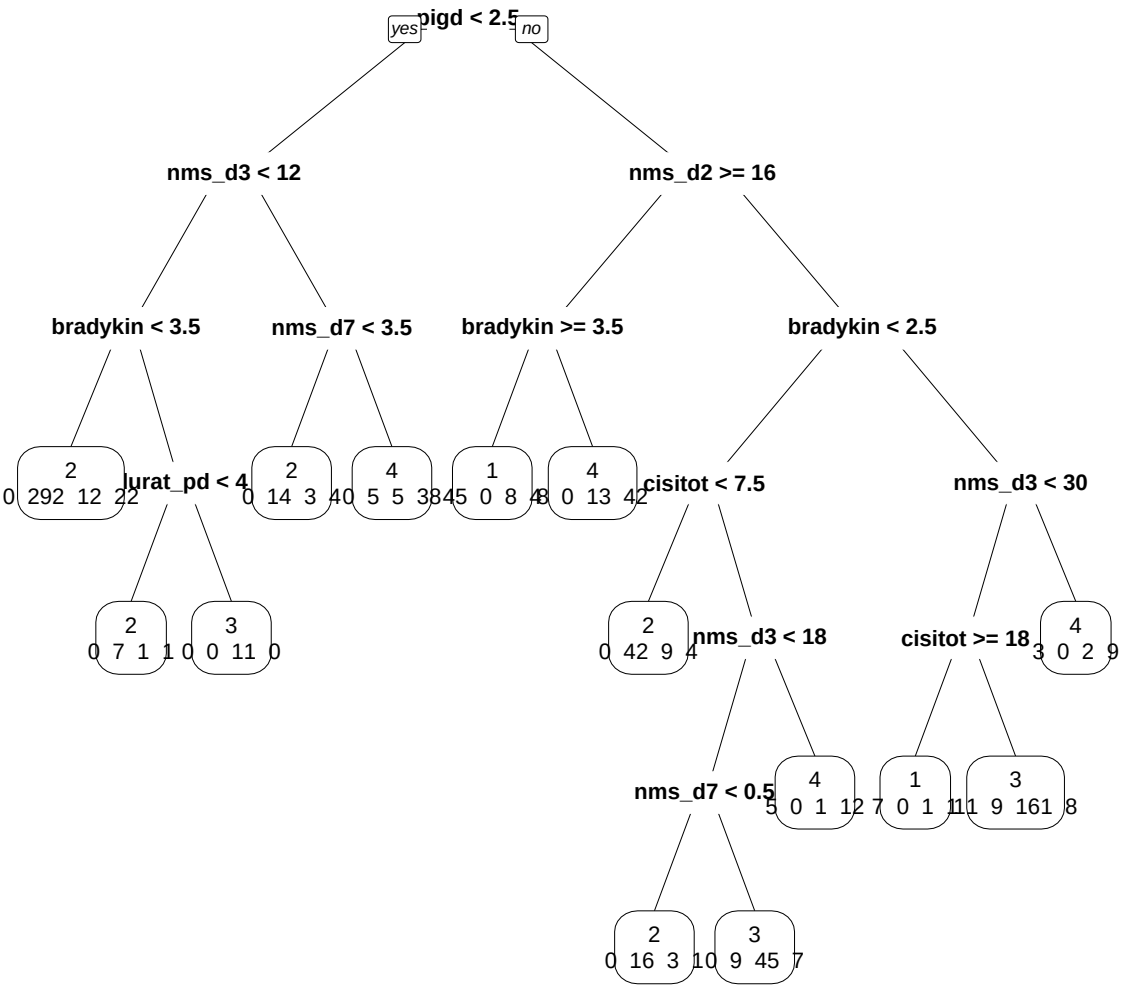

Figure 1: Decision Tree from k-means clustering, 4 clusters

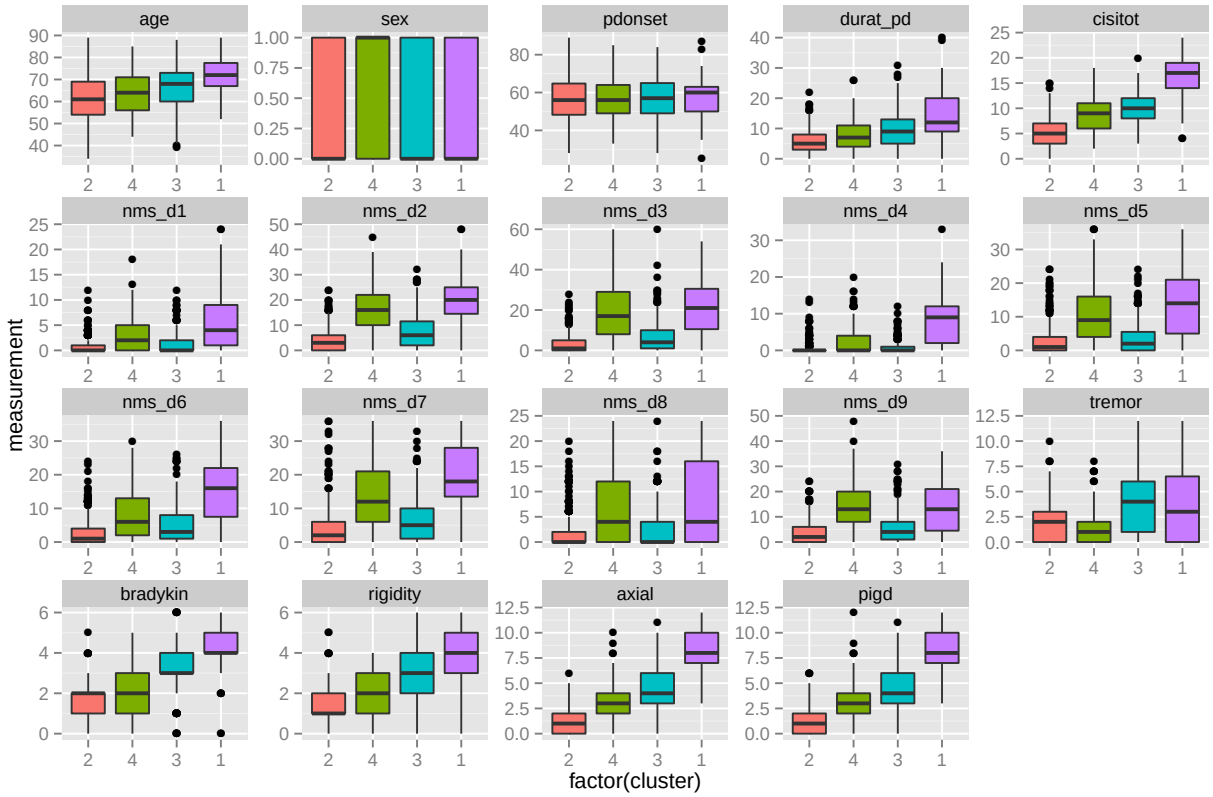

Figure 2: Cluster Summaries, $k = 4$

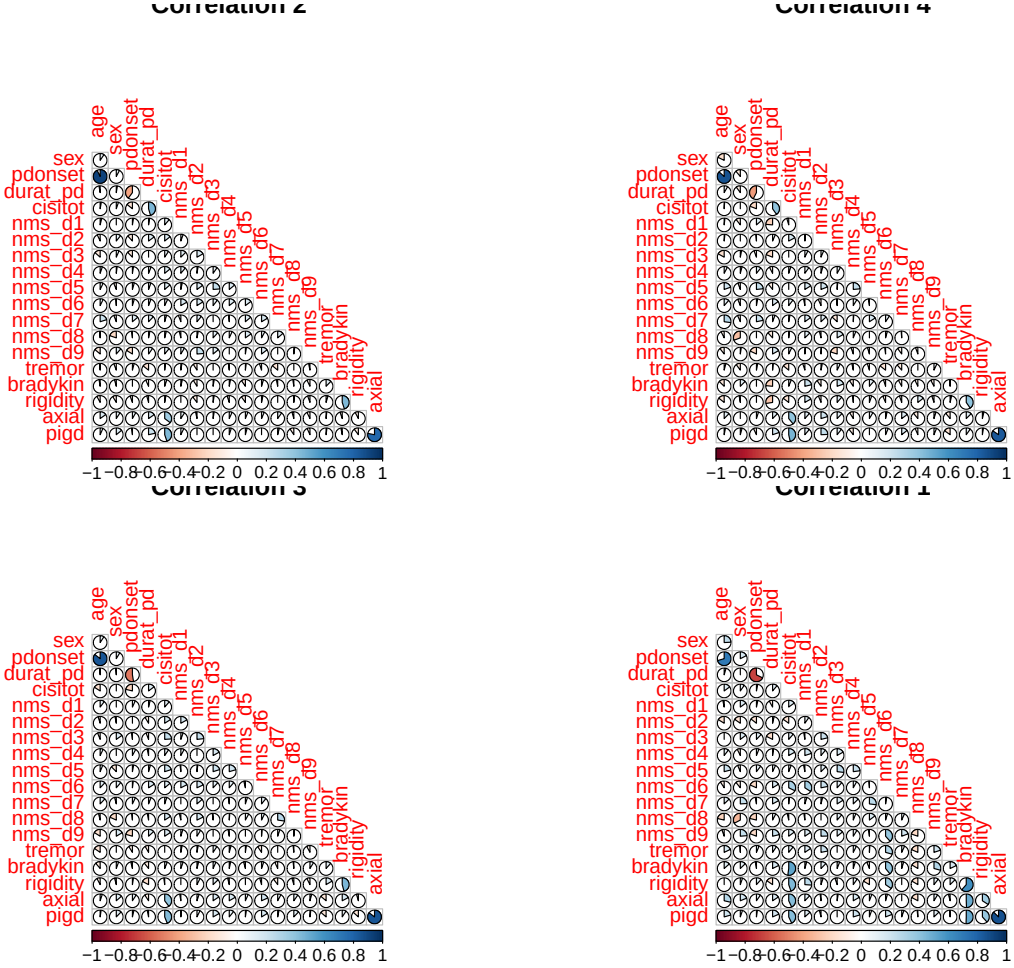

Figure 3: Correlation plots

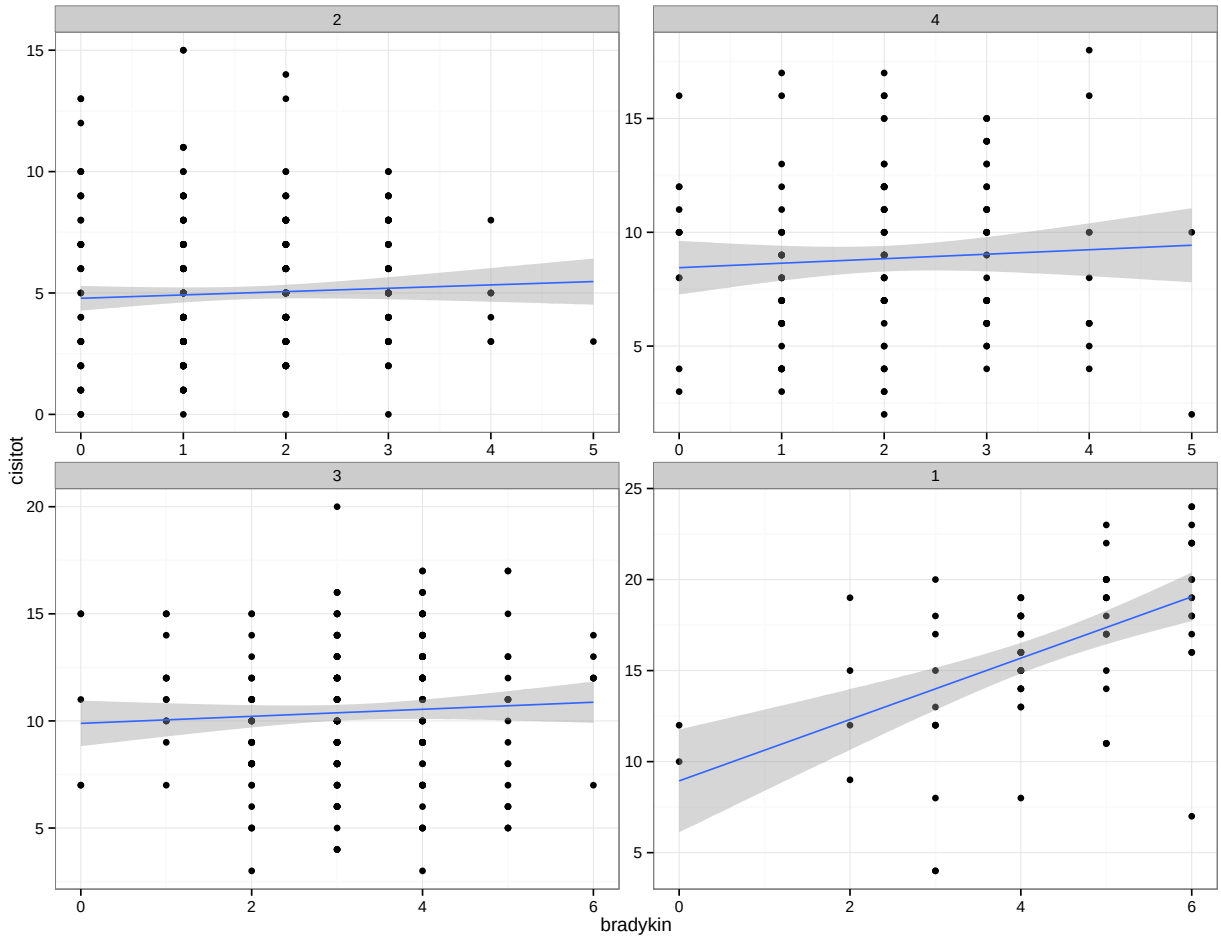

Figure 4: Relationship between bradykinesia and cisitot

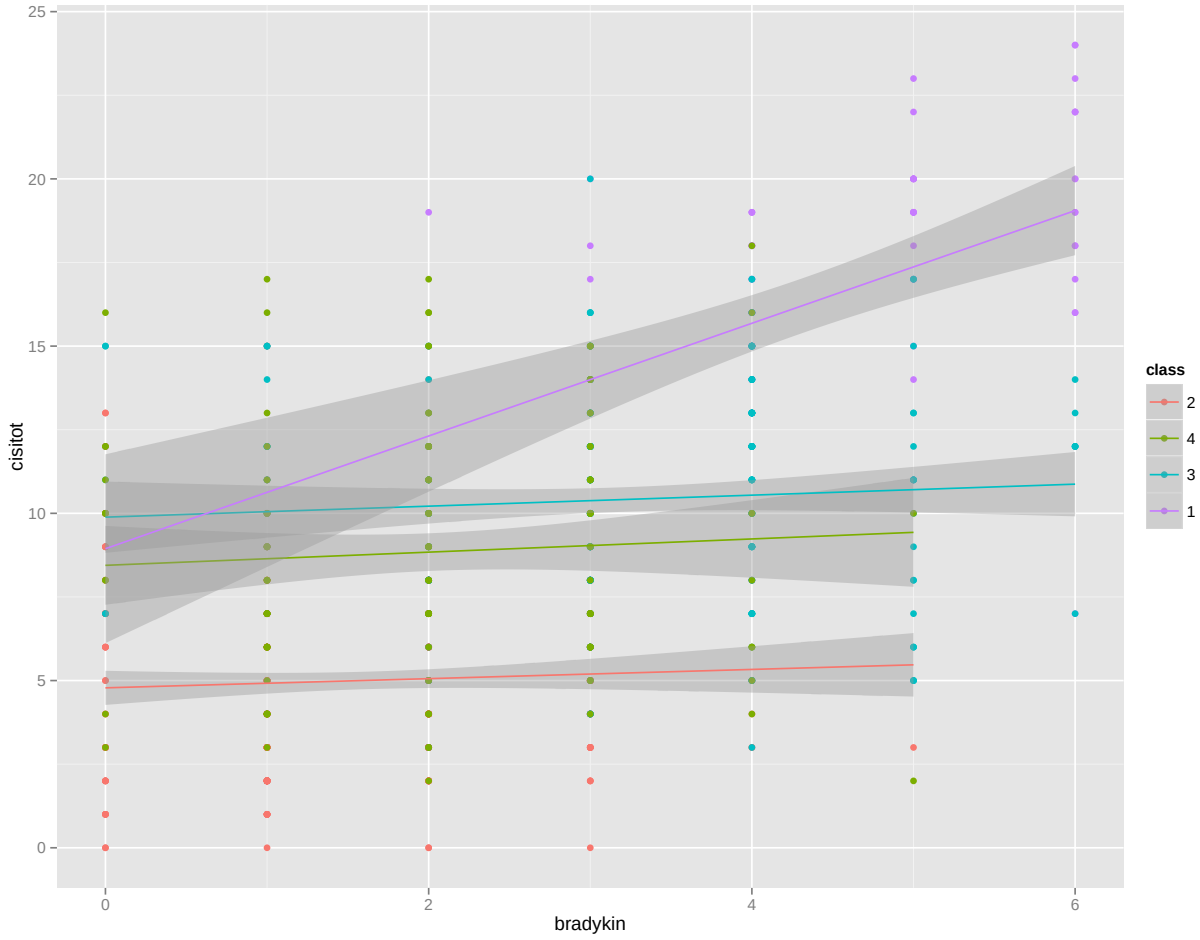

Figure 5: Relationship between bradykinesia and cisitot (condensed)

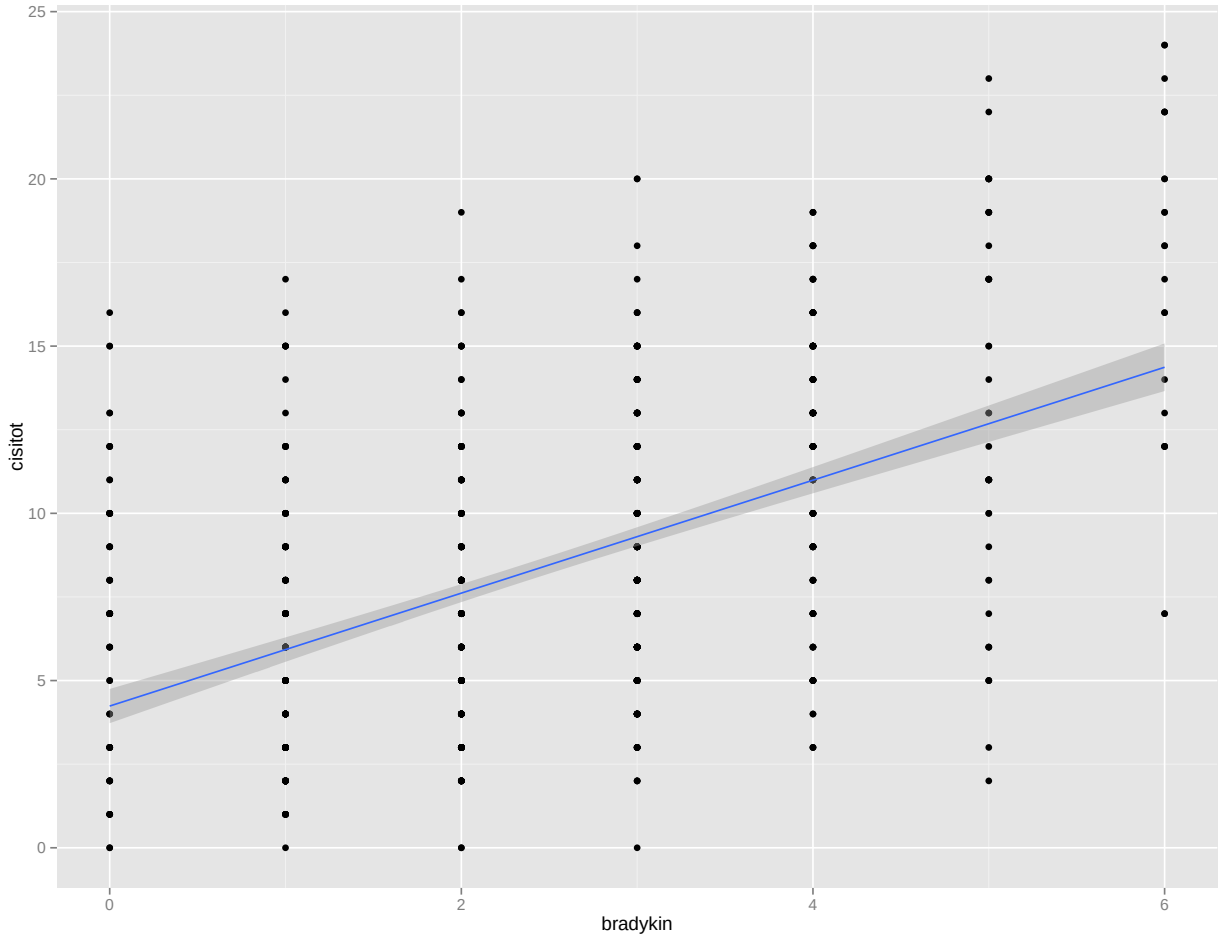

Figure 6: Relationship between bradykinesia and cisitot (entire data)

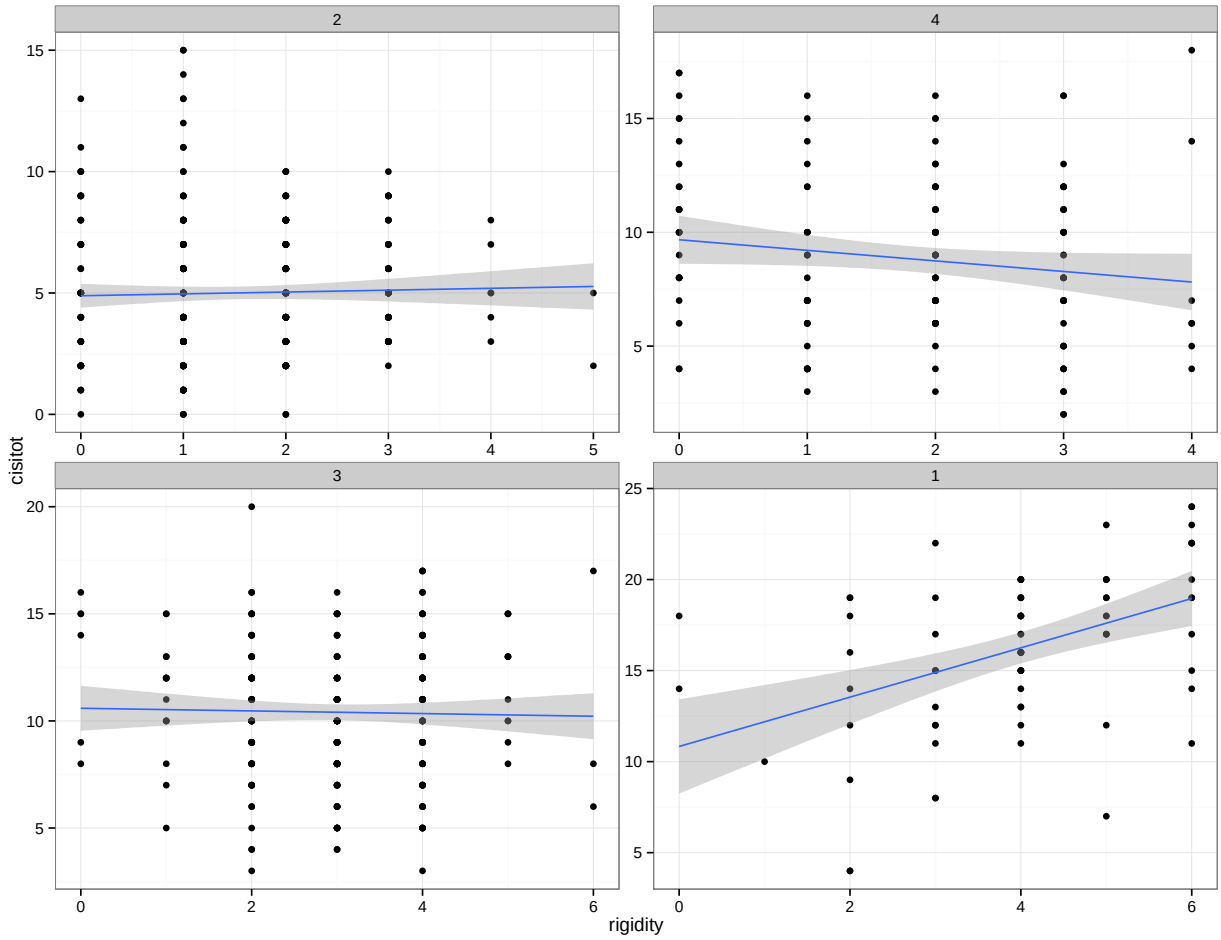

Figure 7: Similar relationship between rigidity and cisitot

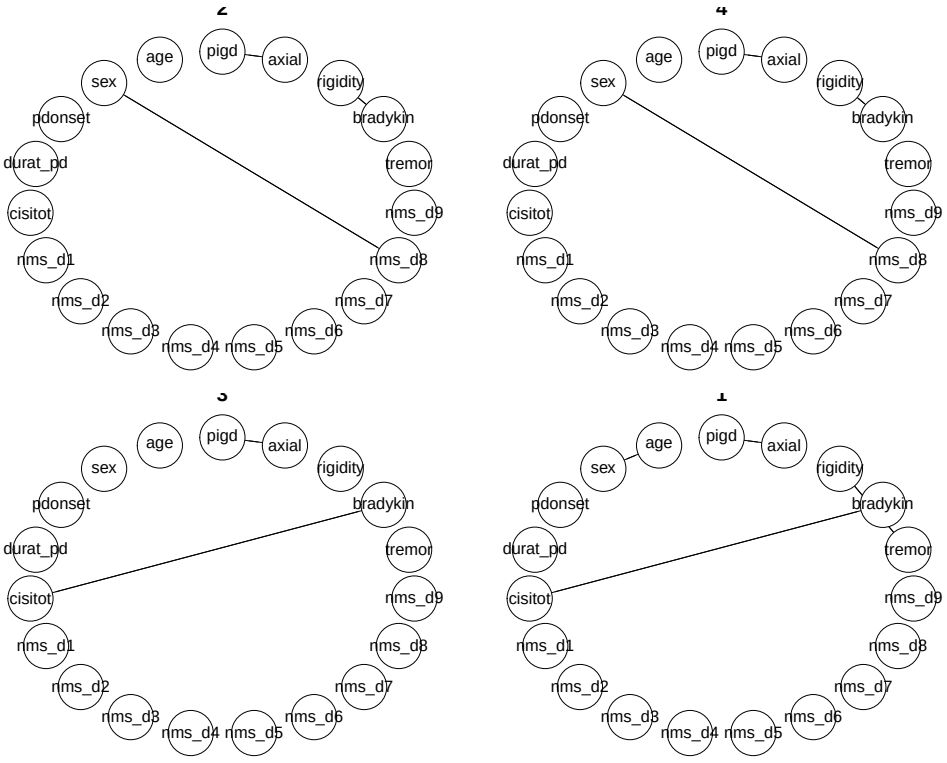

Figure 8: Bayesian Networks