Human Activity Recognition - Clustering Analysis

swirl - Data Science - Exploratory data analysis - Clustering

18.10.2021

1. Intro

The goal is to conduct exploratory data analysis on the Human Activity Recognition Using Smartphones Data Set from the UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

The study creating this database involved 30 volunteers performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. … Each person performed six activities … wearing a smartphone (Samsung Galaxy S II) on the waist. … The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The “getting_cleaning_data.R” file aggregates both the Train and test data sets in a single Data frame. Than creates a tidy dataset, just with the mean ans standard deviation columns.

The “clustering_analysis.R” analyses and plots the data, applying Hierarchical and K-Means clustering methods.

2. Data

Let’s understand tha Data: what kind of data we have?

There are 10299 measurements/observations (train and test data set together) and 564 features/columns in the data set. In addition to the 561 features of the measurements, there are added 3 extra columns.

```
## [1] "Subject"         "Activity_Labels" "Activity"
```

Let’s look at the dataframe

Table 1: The Data (the first 6 columns)

tBodyAcc-mean()-X	tBodyAcc-mean()-Y	tBodyAcc-mean()-Z	tBodyAcc-std()-X	tBodyAcc-std()-Y	tBodyAcc-std()-Z
0.2885845	-0.0202942	-0.1329051	-0.9952786	-0.9831106	-0.9135264
0.2784188	-0.0164106	-0.1235202	-0.9982453	-0.9753002	-0.9603220
0.2796531	-0.0194672	-0.1134617	-0.9953796	-0.9671870	-0.9789440
0.2791739	-0.0262006	-0.1232826	-0.9960915	-0.9834027	-0.9906751
0.2766288	-0.0165697	-0.1153618	-0.9981386	-0.9808173	-0.9904816
0.2771988	-0.0100978	-0.1051372	-0.9973350	-0.9904868	-0.9954200

Summary of the Data

```
##  tBodyAcc-mean()-X tBodyAcc-mean()-Y  tBodyAcc-mean()-Z  tBodyAcc-std()-X
##  Min.   :-1.0000   Min.   :-1.00000   Min.   :-1.00000   Min.   :-1.0000
##  1st Qu.: 0.2626   1st Qu.:-0.02490   1st Qu.:-0.12102   1st Qu.:-0.9924
##  Median : 0.2772   Median :-0.01716   Median :-0.10860   Median :-0.9430
##  Mean   : 0.2743   Mean   :-0.01774   Mean   :-0.10892   Mean   :-0.6078
##  3rd Qu.: 0.2884   3rd Qu.:-0.01062   3rd Qu.:-0.09759   3rd Qu.:-0.2503
##  Max.   : 1.0000   Max.   : 1.00000   Max.   : 1.00000   Max.   : 1.0000
##  tBodyAcc-std()-Y   tBodyAcc-std()-Z
##  Min.   :-1.00000   Min.   :-1.0000
##  1st Qu.:-0.97699   1st Qu.:-0.9791
##  Median :-0.83503   Median :-0.8508
##  Mean   :-0.51019   Mean   :-0.6131
##  3rd Qu.:-0.05734   3rd Qu.:-0.2787
##  Max.   : 1.00000   Max.   : 1.0000
```

Measurements by Subject

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
## 347 302 341 317 302 325 308 281 288 294 316 320 327 323 328 366 368 364 360 354
##  21  22  23  24  25  26  27  28  29  30
## 408 321 372 381 409 392 376 382 344 383
```

Measurements by Activity

```
##
##   laying  sitting standing     walk walkdown   walkup
##     1944     1777     1906     1722     1406     1544
```

3. CodeBook, Variables

```
## ’data.frame’:    10299 obs. of  564 variables:
##  $ tBodyAcc-mean()-X                    : num  0.289 0.278 0.28 0.279 0.277 ...
##  $ tBodyAcc-mean()-Y                    : num  -0.0203 -0.0164 -0.0195 -0.0262 -0.0166 ...
##  $ tBodyAcc-mean()-Z                    : num  -0.133 -0.124 -0.113 -0.123 -0.115 ...
##  $ tBodyAcc-std()-X                     : num  -0.995 -0.998 -0.995 -0.996 -0.998 ...
##  $ tBodyAcc-std()-Y                     : num  -0.983 -0.975 -0.967 -0.983 -0.981 ...
##  $ tBodyAcc-std()-Z                     : num  -0.914 -0.96 -0.979 -0.991 -0.99 ...
##  $ tBodyAcc-mad()-X                     : num  -0.995 -0.999 -0.997 -0.997 -0.998 ...
##  $ tBodyAcc-mad()-Y                     : num  -0.983 -0.975 -0.964 -0.983 -0.98 ...
##  $ tBodyAcc-mad()-Z                     : num  -0.924 -0.958 -0.977 -0.989 -0.99 ...
##  $ tBodyAcc-max()-X                     : num  -0.935 -0.943 -0.939 -0.939 -0.942 ...
##  $ tBodyAcc-max()-Y                     : num  -0.567 -0.558 -0.558 -0.576 -0.569 ...
##  $ tBodyAcc-max()-Z                     : num  -0.744 -0.818 -0.818 -0.83 -0.825 ...
##  $ tBodyAcc-min()-X                     : num  0.853 0.849 0.844 0.844 0.849 ...
##  $ tBodyAcc-min()-Y                     : num  0.686 0.686 0.682 0.682 0.683 ...
##  $ tBodyAcc-min()-Z                     : num  0.814 0.823 0.839 0.838 0.838 ...
##  $ tBodyAcc-sma()                       : num  -0.966 -0.982 -0.983 -0.986 -0.993 ...
##  $ tBodyAcc-energy()-X                  : num  -1 -1 -1 -1 -1 ...
##  $ tBodyAcc-energy()-Y                  : num  -1 -1 -1 -1 -1 ...
##  $ tBodyAcc-energy()-Z                  : num  -0.995 -0.998 -0.999 -1 -1 ...
##  $ tBodyAcc-iqr()-X                     : num  -0.994 -0.999 -0.997 -0.997 -0.998 ...
##  $ tBodyAcc-iqr()-Y                    : num  -0.988 -0.978 -0.965 -0.984 -0.981 ...
##  $ tBodyAcc-iqr()-Z                    : num  -0.943 -0.948 -0.975 -0.986 -0.991 ...
##  $ tBodyAcc-entropy()-X                : num  -0.408 -0.715 -0.592 -0.627 -0.787 ...
##  $ tBodyAcc-entropy()-Y                : num  -0.679 -0.501 -0.486 -0.851 -0.559 ...
##  $ tBodyAcc-entropy()-Z                : num  -0.602 -0.571 -0.571 -0.912 -0.761 ...
##  $ tBodyAcc-arCoeff()-X,1              : num  0.9293 0.6116 0.273 0.0614 0.3133 ...
##  $ tBodyAcc-arCoeff()-X,2              : num  -0.853 -0.3295 -0.0863 0.0748 -0.1312 ...
##  $ tBodyAcc-arCoeff()-X,3              : num  0.36 0.284 0.337 0.198 0.191 ...
##  $ tBodyAcc-arCoeff()-X,4              : num  -0.0585 0.2846 -0.1647 -0.2643 0.0869 ...
##  $ tBodyAcc-arCoeff()-Y,1              : num  0.2569 0.1157 0.0172 0.0725 0.2576 ...
##  $ tBodyAcc-arCoeff()-Y,2              : num  -0.2248 -0.091 -0.0745 -0.1553 -0.2725 ...
##  $ tBodyAcc-arCoeff()-Y,3              : num  0.264 0.294 0.342 0.323 0.435 ...
##  $ tBodyAcc-arCoeff()-Y,4              : num  -0.0952 -0.2812 -0.3326 -0.1708 -0.3154 ...
##  $ tBodyAcc-arCoeff()-Z,1              : num  0.279 0.086 0.239 0.295 0.44 ...
##  $ tBodyAcc-arCoeff()-Z,2              : num  -0.4651 -0.0222 -0.1362 -0.3061 -0.2691 ...
##  $ tBodyAcc-arCoeff()-Z,3              : num  0.4919 -0.0167 0.1739 0.4821 0.1794 ...
##  $ tBodyAcc-arCoeff()-Z,4              : num  -0.191 -0.221 -0.299 -0.47 -0.089 ...
##  $ tBodyAcc-correlation()-X,Y          : num  0.3763 -0.0134 -0.1247 -0.3057 -0.1558 ...
##  $ tBodyAcc-correlation()-X,Z          : num  0.4351 -0.0727 -0.1811 -0.3627 -0.1898 ...
##  $ tBodyAcc-correlation()-Y,Z          : num  0.661 0.579 0.609 0.507 0.599 ...
##  $ tGravityAcc-mean()-X                : num  0.963 0.967 0.967 0.968 0.968 ...
##  $ tGravityAcc-mean()-Y                : num  -0.141 -0.142 -0.142 -0.144 -0.149 ...
##  $ tGravityAcc-mean()-Z                : num  0.1154 0.1094 0.1019 0.0999 0.0945 ...
##  $ tGravityAcc-std()-X                 : num  -0.985 -0.997 -1 -0.997 -0.998 ...
##  $ tGravityAcc-std()-Y                 : num  -0.982 -0.989 -0.993 -0.981 -0.988 ...
##  $ tGravityAcc-std()-Z                 : num  -0.878 -0.932 -0.993 -0.978 -0.979 ...
##  $ tGravityAcc-mad()-X                 : num  -0.985 -0.998 -1 -0.996 -0.998 ...
##  $ tGravityAcc-mad()-Y                 : num  -0.984 -0.99 -0.993 -0.981 -0.989 ...
##  $ tGravityAcc-mad()-Z                 : num  -0.895 -0.933 -0.993 -0.978 -0.979 ...
##  $ tGravityAcc-max()-X                 : num  0.892 0.892 0.892 0.894 0.894 ...
##  $ tGravityAcc-max()-Y                 : num  -0.161 -0.161 -0.164 -0.164 -0.167 ...
##  $ tGravityAcc-max()-Z                 : num  0.1247 0.1226 0.0946 0.0934 0.0917 ...
##  $ tGravityAcc-min()-X                 : num  0.977 0.985 0.987 0.987 0.987 ...
##  $ tGravityAcc-min()-Y                 : num  -0.123 -0.115 -0.115 -0.121 -0.122 ...
##  $ tGravityAcc-min()-Z                 : num  0.0565 0.1028 0.1028 0.0958 0.0941 ...
##  $ tGravityAcc-sma()                   : num  -0.375 -0.383 -0.402 -0.4 -0.4 ...
##  $ tGravityAcc-energy()-X              : num  0.899 0.908 0.909 0.911 0.912 ...
##  $ tGravityAcc-energy()-Y              : num  -0.971 -0.971 -0.97 -0.969 -0.967 ...
##  $ tGravityAcc-energy()-Z              : num  -0.976 -0.979 -0.982 -0.982 -0.984 ...
##  $ tGravityAcc-iqr()-X                 : num  -0.984 -0.999 -1 -0.996 -0.998 ...
##  $ tGravityAcc-iqr()-Y                 : num  -0.989 -0.99 -0.992 -0.981 -0.991 ...
##  $ tGravityAcc-iqr()-Z                 : num  -0.918 -0.942 -0.993 -0.98 -0.98 ...
##  $ tGravityAcc-entropy()-X             : num  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ tGravityAcc-entropy()-Y             : num  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ tGravityAcc-entropy()-Z             : num  0.114 -0.21 -0.927 -0.596 -0.617 ...
##  $ tGravityAcc-arCoeff()-X,1           : num  -0.59042 -0.41006 0.00223 -0.06493 -0.25727 ...
##  $ tGravityAcc-arCoeff()-X,2           : num  0.5911 0.4139 0.0275 0.0754 0.2689 ...
##  $ tGravityAcc-arCoeff()-X,3           : num  -0.5918 -0.4176 -0.0567 -0.0858 -0.2807 ...
##  $ tGravityAcc-arCoeff()-X,4           : num  0.5925 0.4213 0.0855 0.0962 0.2926 ...
##  $ tGravityAcc-arCoeff()-Y,1           : num  -0.745 -0.196 -0.329 -0.295 -0.167 ...
##  $ tGravityAcc-arCoeff()-Y,2           : num  0.7209 0.1253 0.2705 0.2283 0.0899 ...
##  $ tGravityAcc-arCoeff()-Y,3           : num  -0.7124 -0.1056 -0.2545 -0.2063 -0.0663 ...
##  $ tGravityAcc-arCoeff()-Y,4           : num  0.7113 0.1091 0.2576 0.2048 0.0671 ...
##  $ tGravityAcc-arCoeff()-Z,1           : num  -0.995 -0.834 -0.705 -0.385 -0.237 ...
##  $ tGravityAcc-arCoeff()-Z,2           : num  0.996 0.834 0.714 0.386 0.239 ...
##  $ tGravityAcc-arCoeff()-Z,3           : num  -0.996 -0.834 -0.723 -0.387 -0.241 ...
##  $ tGravityAcc-arCoeff()-Z,4           : num  0.992 0.83 0.729 0.385 0.241 ...
##  $ tGravityAcc-correlation()-X,Y       : num  0.57 -0.831 -0.181 -0.991 -0.408 ...
##  $ tGravityAcc-correlation()-X,Z       : num  0.439 -0.866 0.338 -0.969 -0.185 ...
##  $ tGravityAcc-correlation()-Y,Z       : num  0.987 0.974 0.643 0.984 0.965 ...
##  $ tBodyAccJerk-mean()-X               : num  0.078 0.074 0.0736 0.0773 0.0734 ...
##  $ tBodyAccJerk-mean()-Y               : num  0.005 0.00577 0.0031 0.02006 0.01912 ...
##  $ tBodyAccJerk-mean()-Z               : num  -0.06783 0.02938 -0.00905 -0.00986 0.01678 ...
##  $ tBodyAccJerk-std()-X                : num  -0.994 -0.996 -0.991 -0.993 -0.996 ...
##  $ tBodyAccJerk-std()-Y                : num  -0.988 -0.981 -0.981 -0.988 -0.988 ...
##  $ tBodyAccJerk-std()-Z                : num  -0.994 -0.992 -0.99 -0.993 -0.992 ...
##  $ tBodyAccJerk-mad()-X                : num  -0.994 -0.996 -0.991 -0.994 -0.997 ...
##  $ tBodyAccJerk-mad()-Y                : num  -0.986 -0.979 -0.979 -0.986 -0.987 ...
##  $ tBodyAccJerk-mad()-Z                : num  -0.993 -0.991 -0.987 -0.991 -0.991 ...
##  $ tBodyAccJerk-max()-X                : num  -0.985 -0.995 -0.987 -0.987 -0.997 ...
##  $ tBodyAccJerk-max()-Y                : num  -0.992 -0.979 -0.979 -0.992 -0.992 ...
##  $ tBodyAccJerk-max()-Z                : num  -0.993 -0.992 -0.992 -0.99 -0.99 ...
##  $ tBodyAccJerk-min()-X                : num  0.99 0.993 0.988 0.988 0.994 ...
##  $ tBodyAccJerk-min()-Y                : num  0.992 0.992 0.992 0.993 0.993 ...
##  $ tBodyAccJerk-min()-Z                : num  0.991 0.989 0.989 0.993 0.986 ...
##  $ tBodyAccJerk-sma()                  : num  -0.994 -0.991 -0.988 -0.993 -0.994 ...
##  $ tBodyAccJerk-energy()-X             : num  -1 -1 -1 -1 -1 ...
##  $ tBodyAccJerk-energy()-Y             : num  -1 -1 -1 -1 -1 ...
##  $ tBodyAccJerk-energy()-Z             : num  -1 -1 -1 -1 -1 ...
##   [list output truncated]
```

Table: Variables

|| || || ||

4. Exploratory data analysis

Let's look at only the first subject (numbered 1) which has a number of 347 measurements as seen before. See the measurement of the first 6 subjects out of 30:

```
##
##   1   2   3   4   5   6
## 347 302 341 317 302 325
```

Let's do some comparisons of activities now by looking at plots of mean body acceleration in the X and Y directions for the first subject.

We see (Figure 1) that the active activities related to walking (shown in the two blues and magenta) show more variability than the passive activities (shown in black, red, and green), particularly in the X dimension.

Let's see (Figure 2) the same figure but now for all the 30 subjects.

4.1 Hierarchical clustering

Let's try hierarchical clustering to see if we can distinguish the activities more.

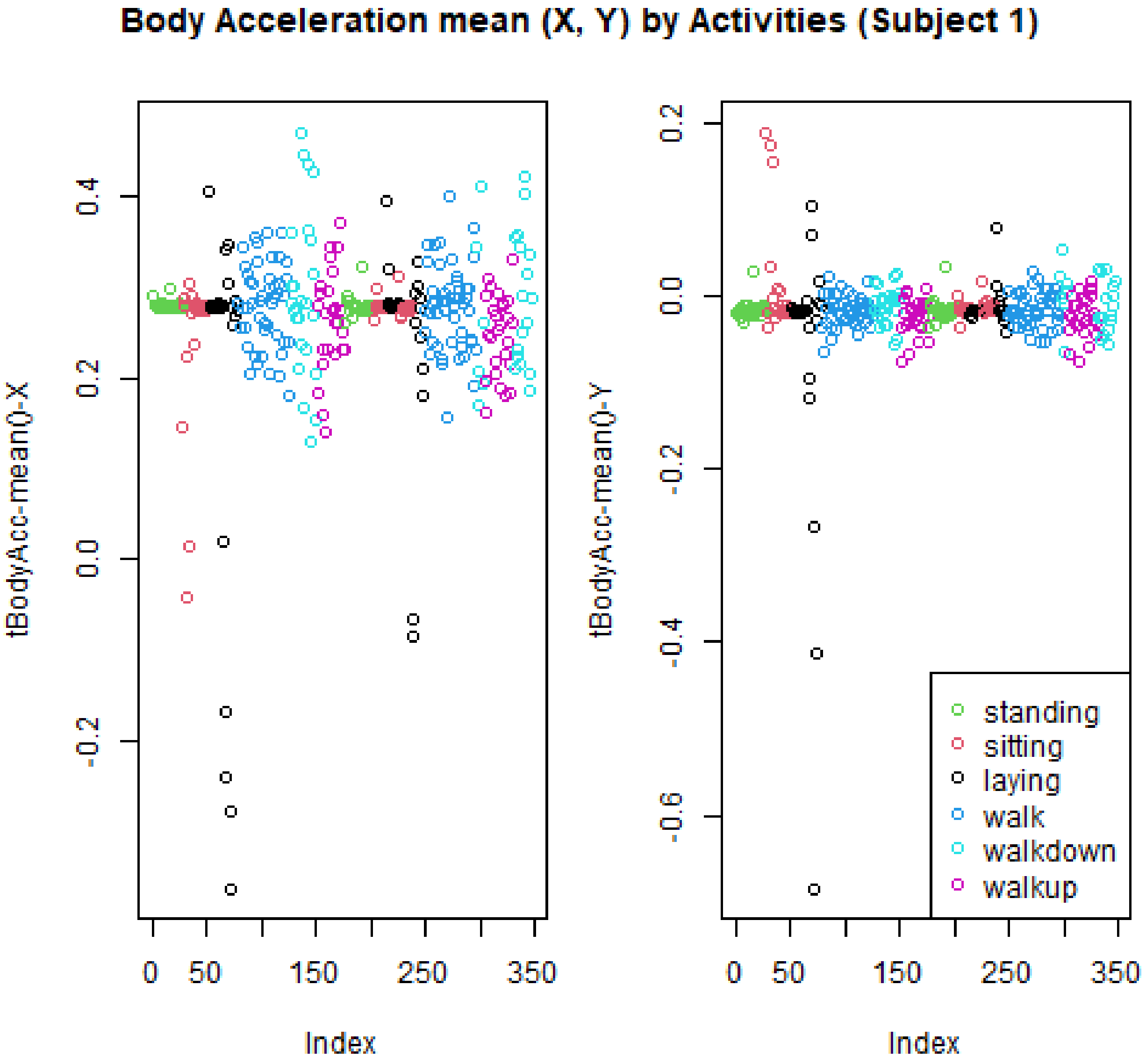

Figure 1: Body Acceleration Mean (X,Y) by Activities (Subject 1)

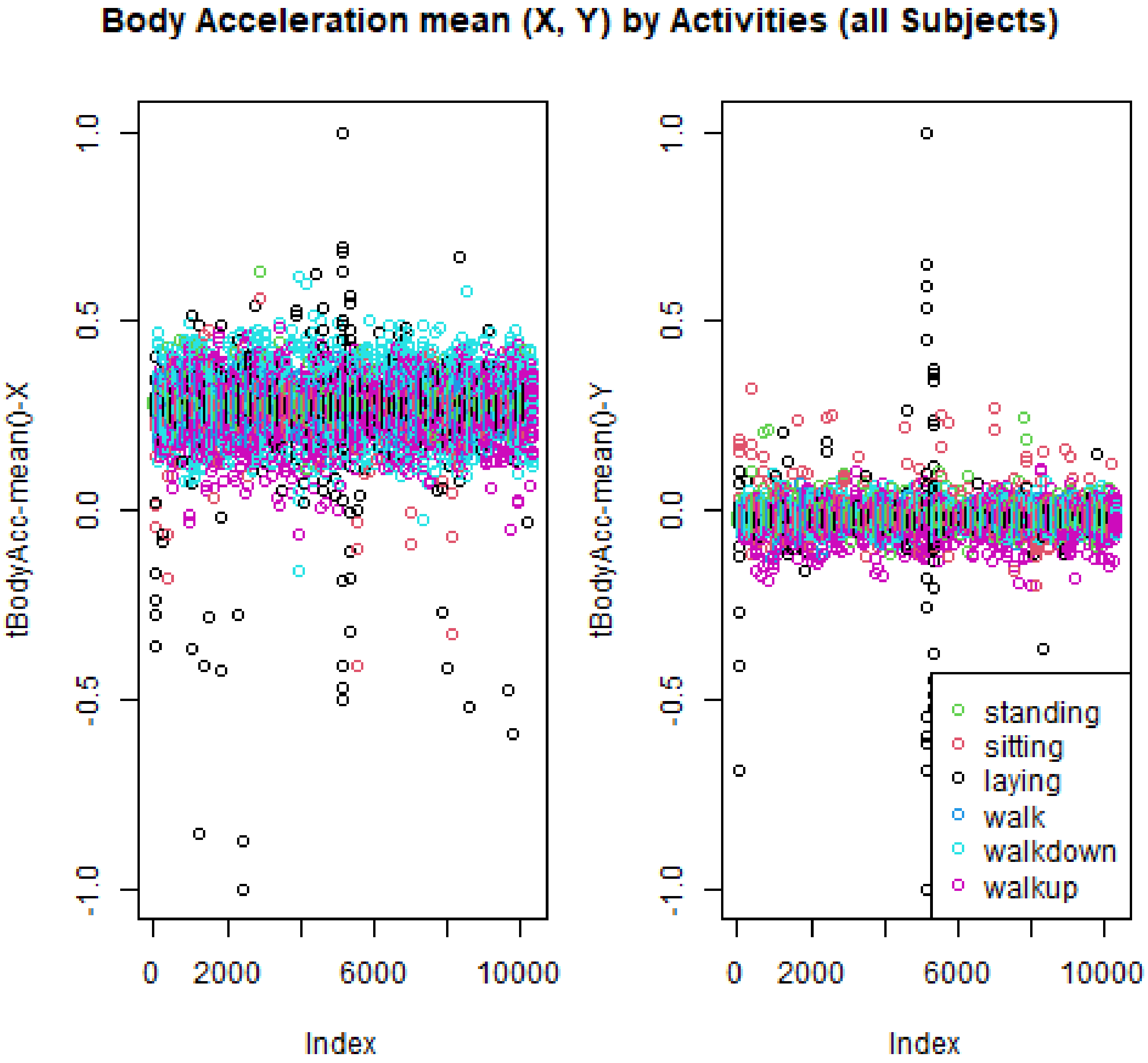

Figure 2: Body Acceleration Mean (X,Y) by Activities (all Subjects)

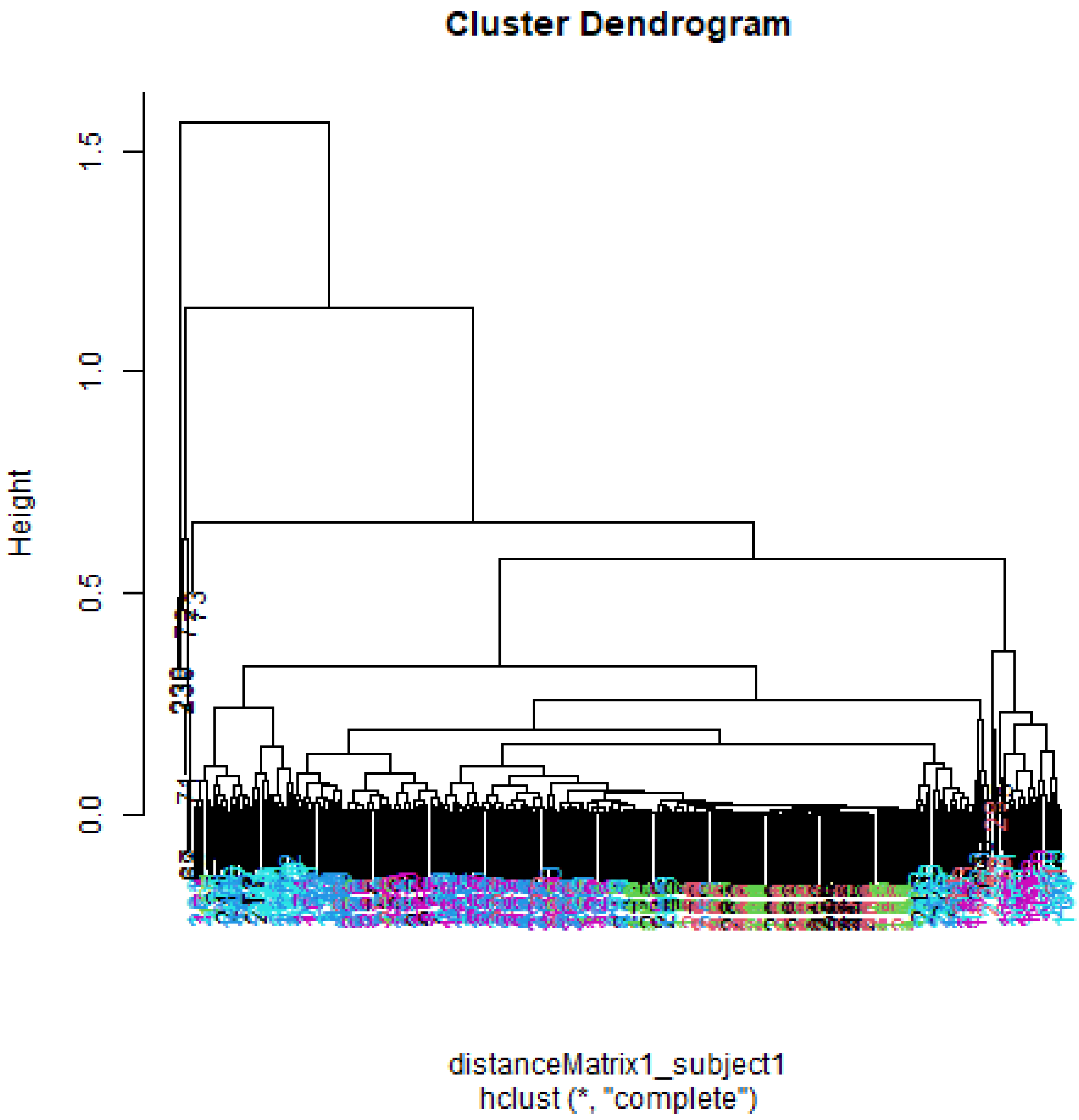

Figure 3: Hierarchical Clustering, active and passive activity colors (Subject 1)

There's no clear grouping of colors (Figure 3), except that active colors (blues and magenta) are near each other as are the passive (black, red, and green) maximum acceleration.

Let's focus then on the 3 dimensions of maximum acceleration.

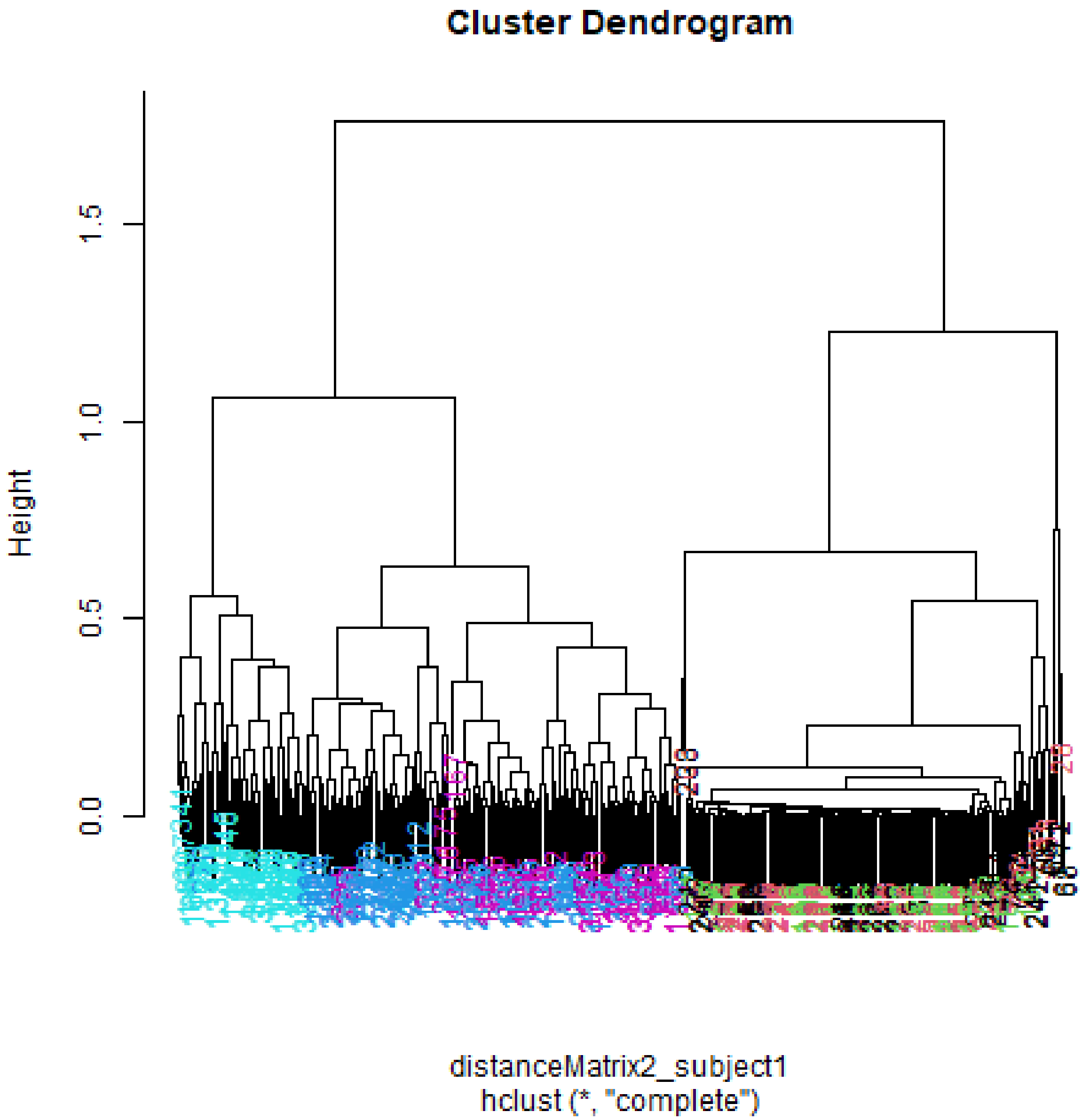

Figure 4: Hierarchical Clustering, 2 Clusters for active and passive activities (Subject 1)

Now we see clearly that the data splits into 2 clusters (Figure 4), active and passive activities. Moreover, the light blue (walking down) is clearly distinct from the other walking activities. The dark blue (walking level) also seems to be somewhat clustered. The passive activities, however, seem all jumbled together with no clear pattern visible.

4.2 SVD Clustering - Singular Value Decomposition

Let's try some SVD now.

4.2.1 Left Singular Vectors, U-Matrix of SVD

Just subject 1:

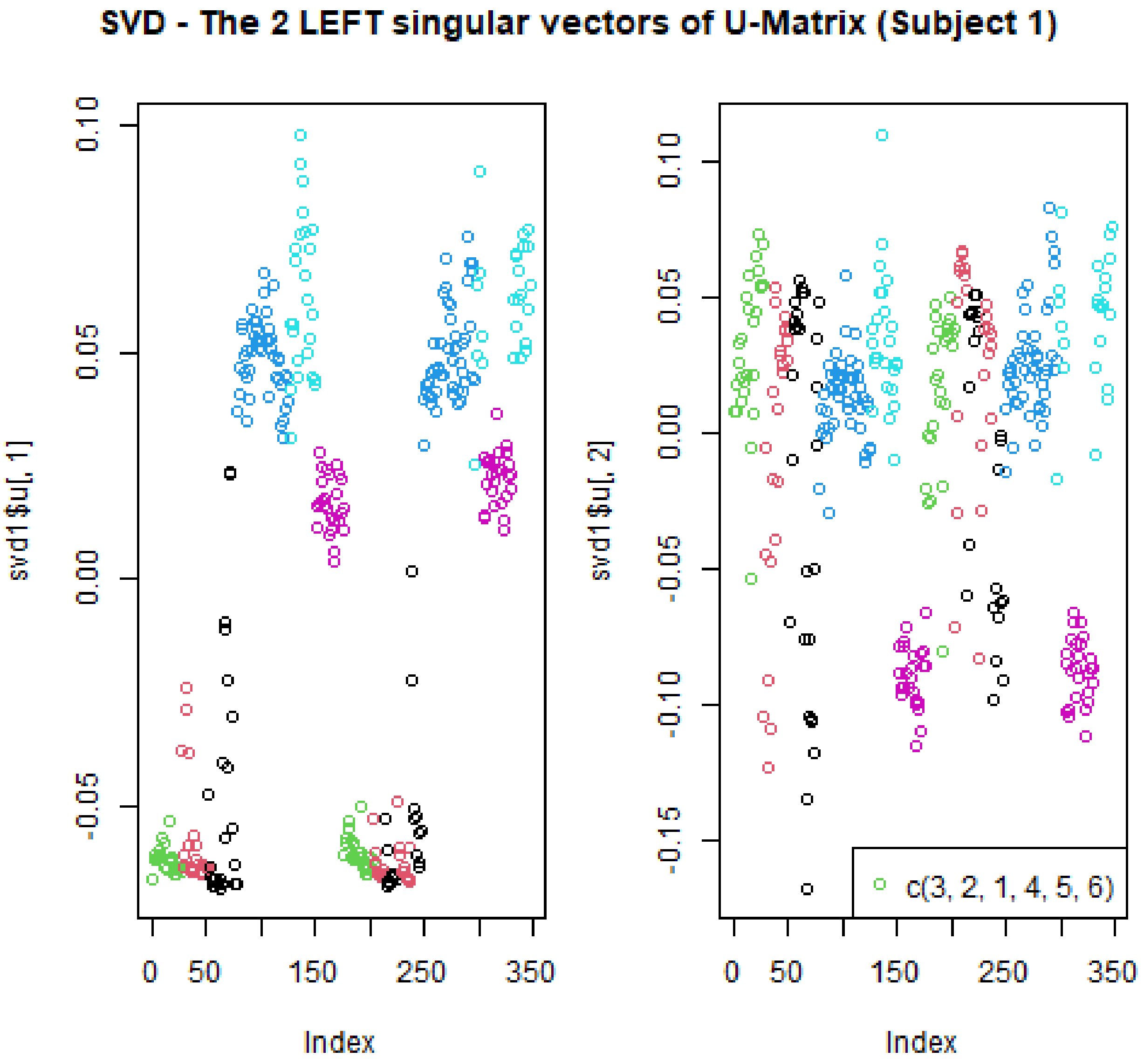

Figure 5: SVD - The 2 LEFT singular vectors of U-Matrix (Subject 1)

All 30 Subjects:

Here (Figure 5) we're looking at the 2 left singular vectors of SVD (the first 2 columns of the U-Matrix of SVD). Each entry of the columns belongs to a particular row with one of the 6 activities assigned to it.

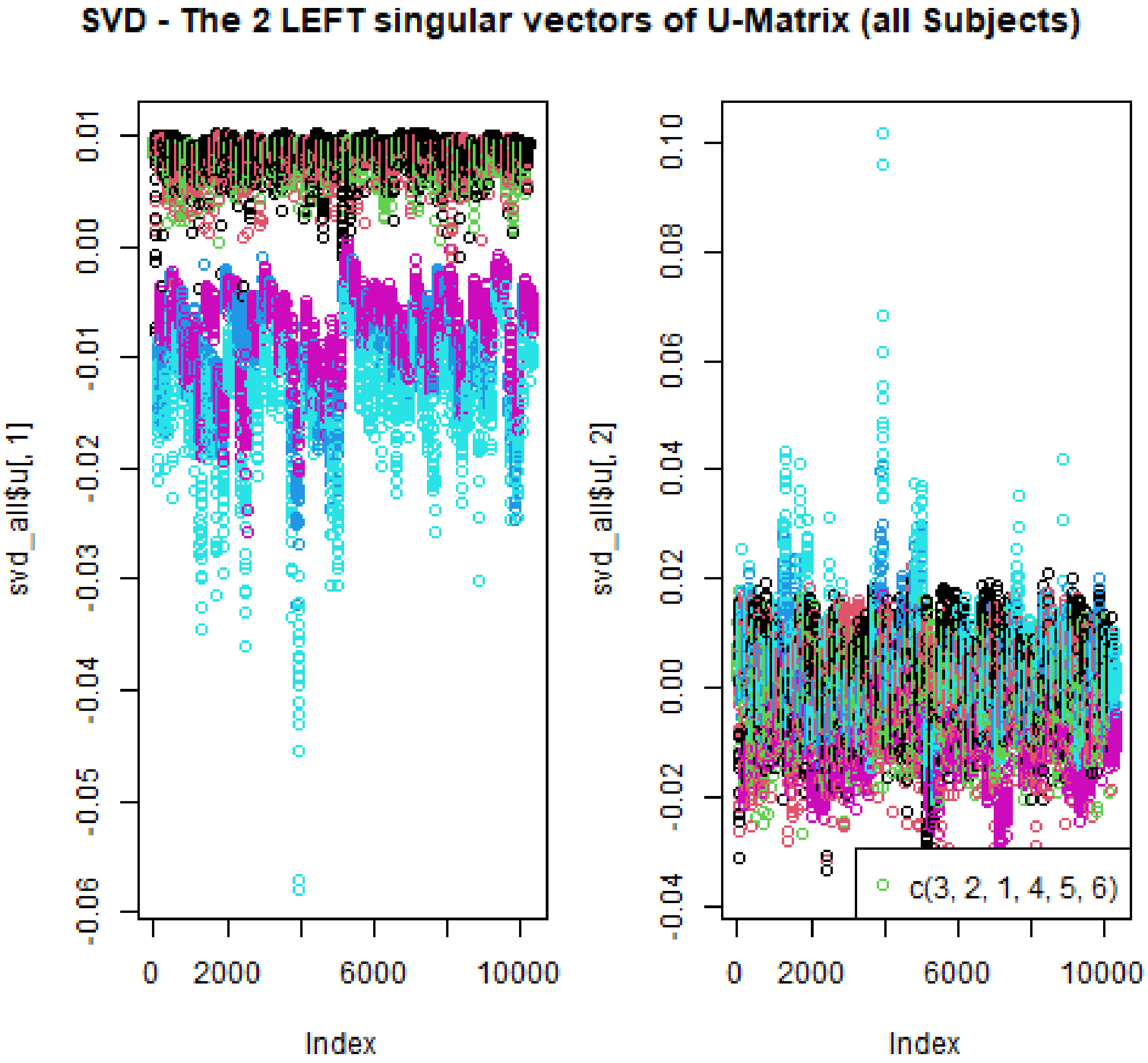

Figure 6: SVD - The 2 LEFT singular vectors of U-Matrix (all Subjects)

We see the activities distinguished by color. Moving from left to right, the first section of rows are green (standing), the second red (sitting), the third black (laying), etc. The first column of u shows separation of the nonmoving (black, red, and green) from the walking activities. The second column is harder to interpret. However, the magenta cluster, which represents walking up, seems separate from the others.

We'll try to figure out why that is. To do that we'll have to find which of the 500+ measurements contributes to the variation of that component. We'll look at the RIGHT singular vectors the columns of V-Matrix), and in particular, the second one since the separation of the magenta cluster stood out in the second column of V-Matrix.

4.2.2 Right Singular Vectors, V-Matrix of SVD

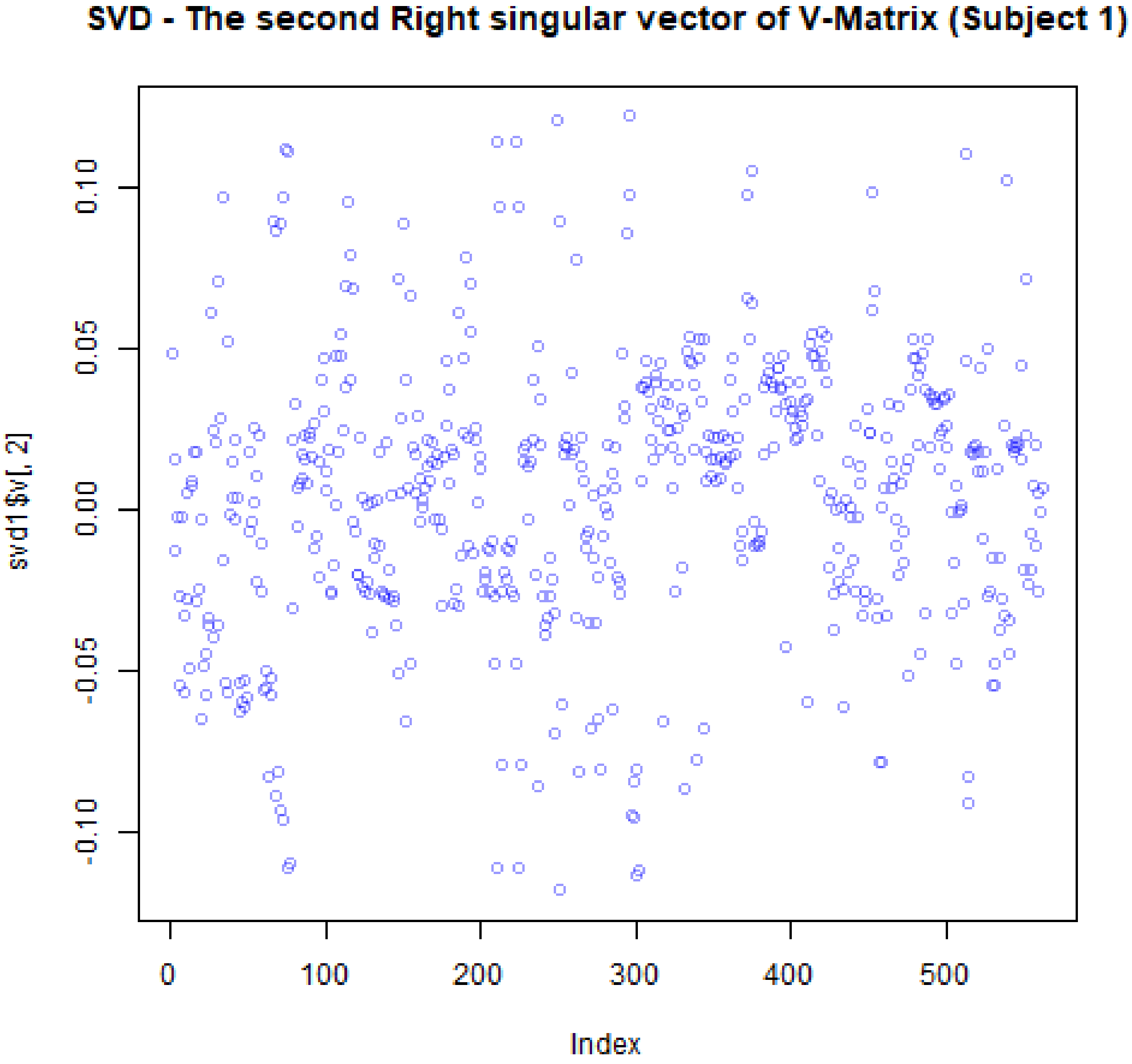

Figure 7: SVD - The second Right singular vector of V-Matrix (Subject 1)

Here's a plot (Figure 7) of the second column of the V-Matrix. We used transparency in our plotting but nothing clearly stands out here. Let's use clustering to find the feature (out of the 500+) which contributes the most to the variation of this second column of the V-Matrix.

Subject 1:

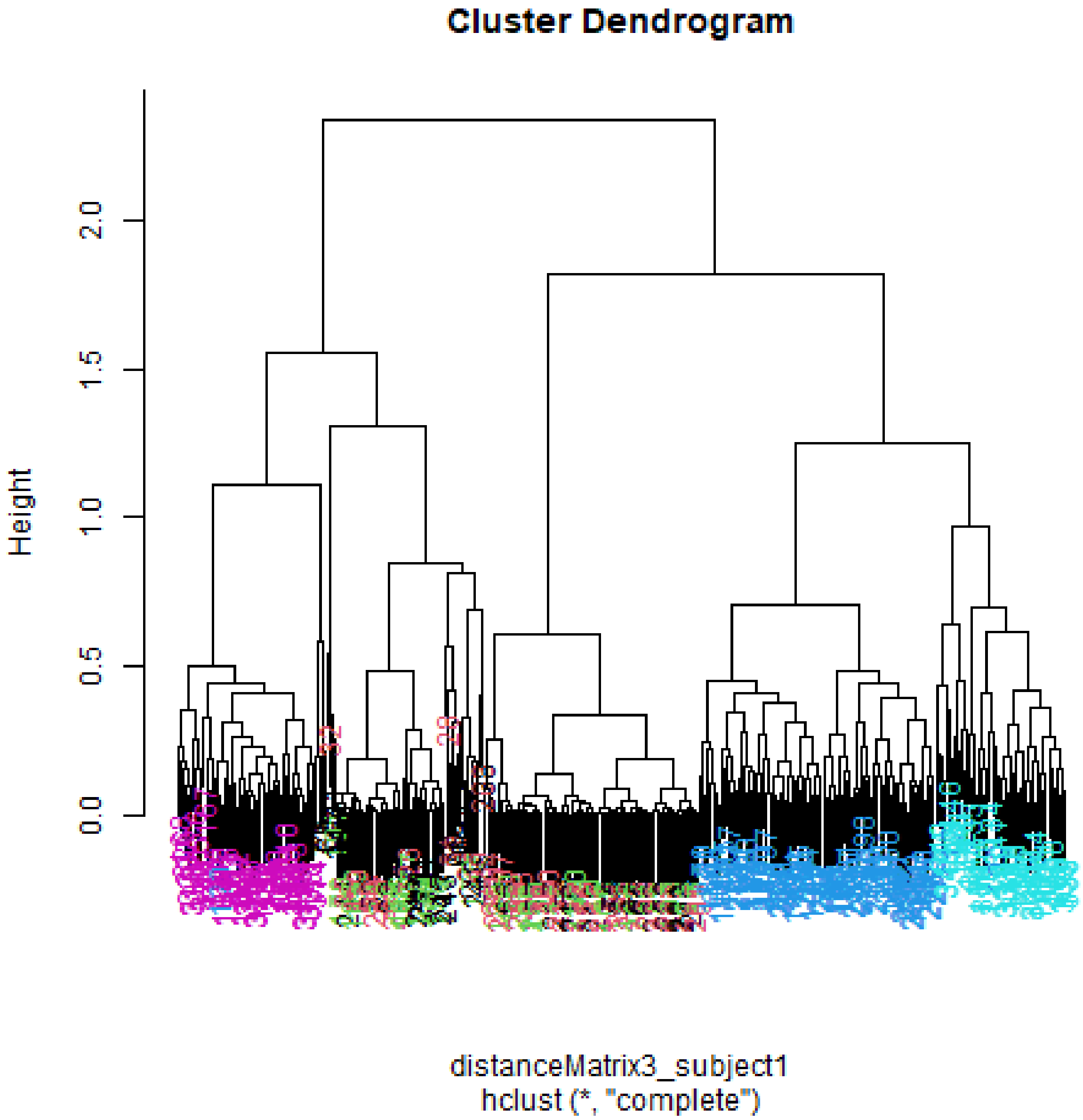

Figure 8: Hierarchical Clustering, Feature Maximal Variationof the second column of V-Matrix (Subject 1)

All Subjects:

Now (Figure 8) we see some real separation. Magenta (walking up) is on the far left, and the two other walking activities, the two blues, are on the far right, but in separate clusters from one another. The nonmoving activities still are jumbled together.

Let's see what measurement is associated with this maximum contributor.

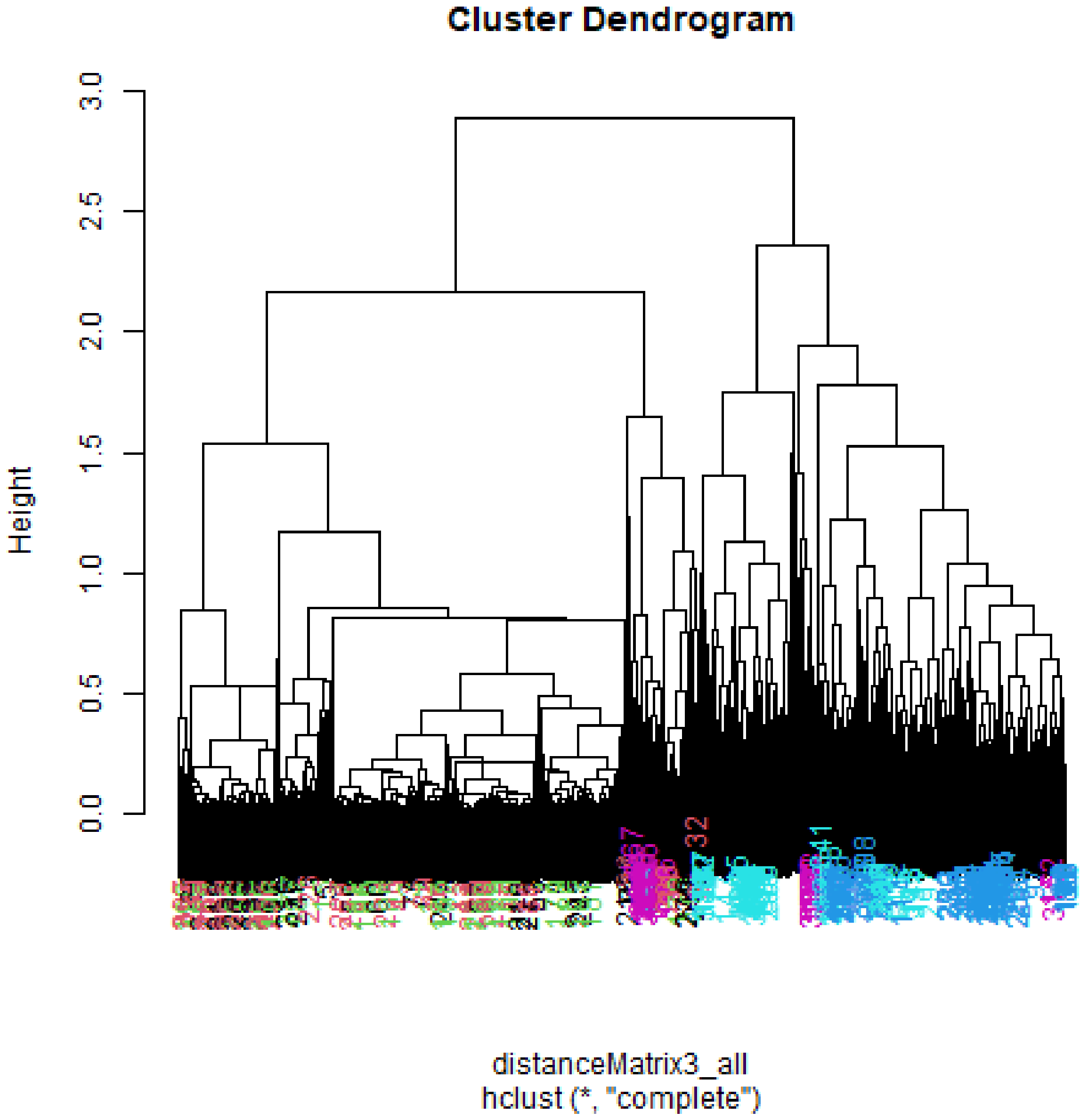

Figure 9: Hierarchical Clustering, Feature Maximal Variationof the second column of V-Matrix (all Subjects)

```
## [1] "fBodyAcc-meanFreq()-Z"
```

So the mean body acceleration in the frequency domain in the Z direction is the main contributor to this clustering phenomenon we're seeing.

4.3 K-Means Clustering

Let's move on to k-means clustering to see if this technique can distinguish between the activities.

Measurements by cluster and activity

```
##
##     laying sitting standing walk walkdown walkup
##   1      9       2        0    0        0      0
##   2     27       0        0    0        0      0
##   3     14      11        3    0        0      0
##   4      0       0        0    0        0     53
##   5      0       0        0   95       49      0
##   6      0      34       50    0        0      0
```

The exact output will depend on the state of the random number generator. The walking activities seem to cluster individually by themselves. This was K-Means with one random start. Now we try K-Means with more random starts and try to return the best one.

```
##
##     laying sitting standing walk walkdown walkup
##   1      0      40       51    0        0      0
##   2      0       0        0    0       49      0
##   3     20       7        2    0        0      0
##   4     29       0        0    0        0      0
##   5      1       0        0    0        0     53
##   6      0       0        0   95        0      0
```

We see that even with 100 random starts, the passive activities tend to cluster together. One of the clusters contains only laying, but in another cluster, standing and sitting group together.

The centers are a 6 by 561 array. Sometimes it's a good idea to look at the features (columns) of these centers to see if any dominate.

We see (Figure 10) the first 3 columns dominate this cluster center. Their names are:

```
## [1] "tBodyAcc-mean()-X" "tBodyAcc-mean()-Y" "tBodyAcc-mean()-Z"
```

So the 3 directions of mean body acceleration seem to have the biggest effect on laying.

See the columns which dominate walkdown.

We see (Figure 11) an interesting pattern here. From left to right, looking at the 12 acceleration measurements in groups of 3, the points decrease in value. The X direction dominates, followed by Y then Z. This might tell us something more about the walking down activity.

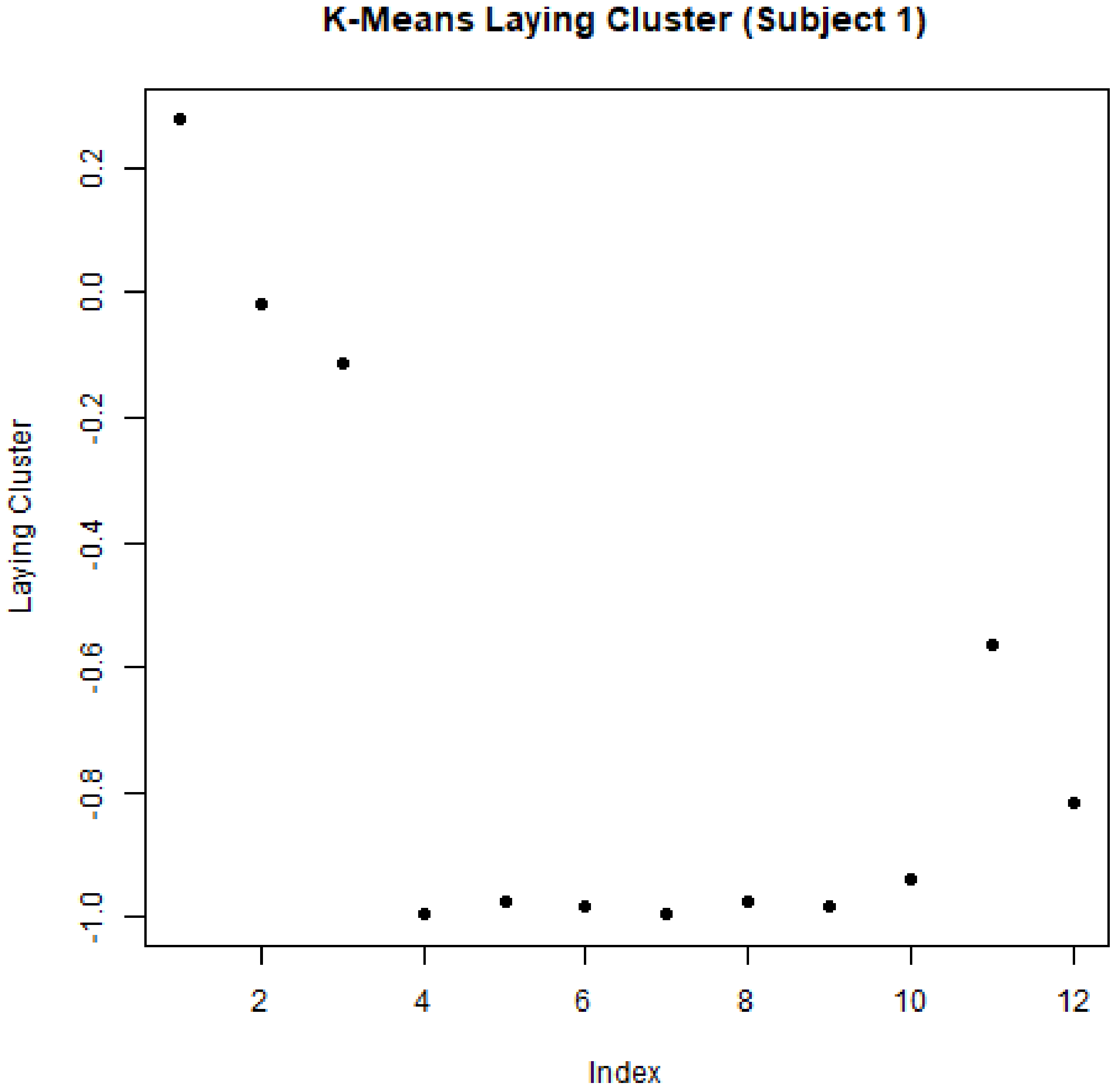

Figure 10: K-Means Laying Cluster (Subject 1)

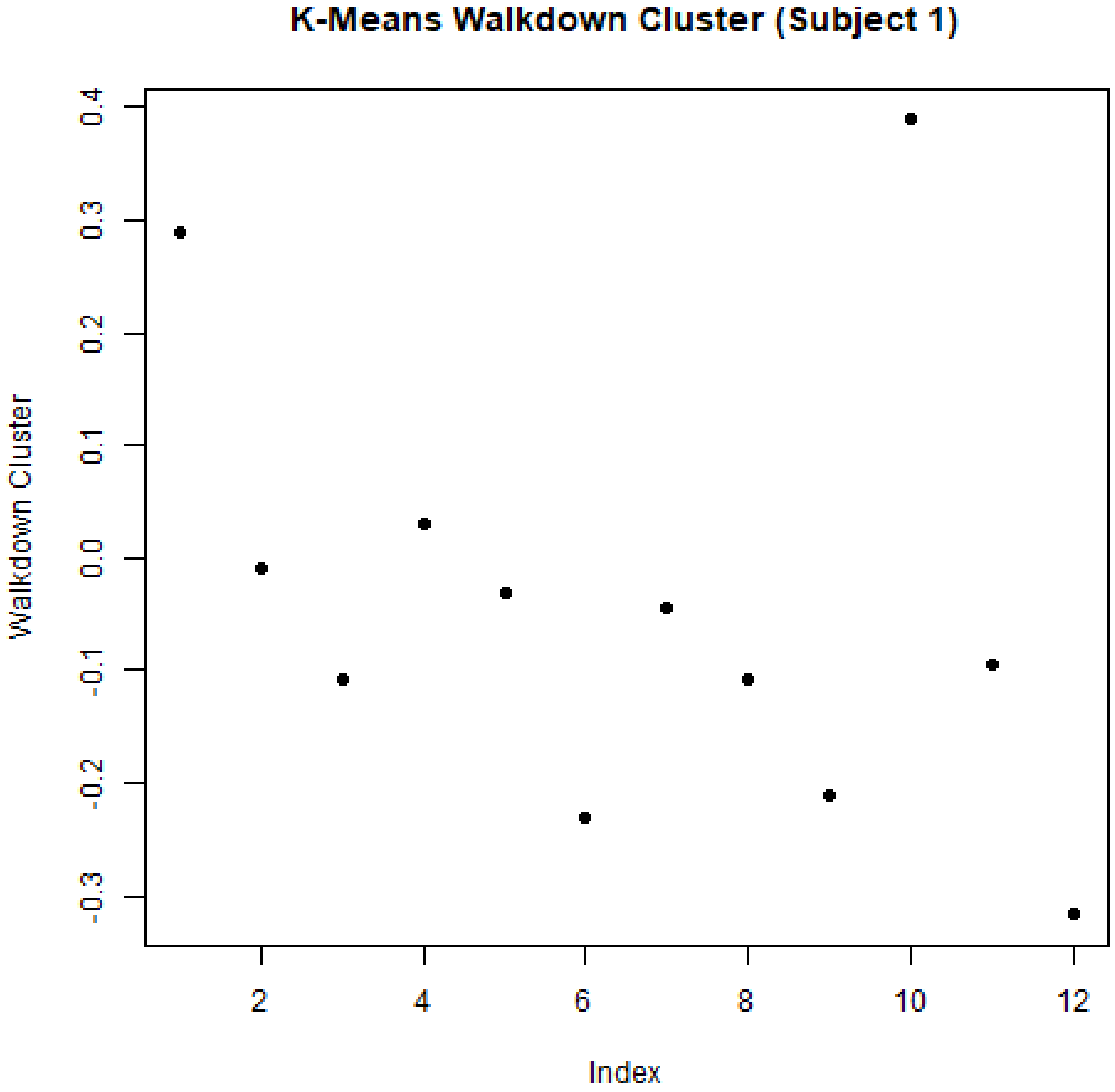

Figure 11: K-Means Walkdown Cluster (Subject 1)

5. Conclusion

We saw here that the sensor measurements were pretty good at discriminating between the 3 walking activities, but the passive activities were harder to distinguish from one another. These might require more analysis or an entirely different set of sensory measurements.

6. Reference

See Johns Hopkins University, Data Science Specialization, Course Exploratory Data Analysis, swirl() exercises on Clustering.