Try PCA for embarc

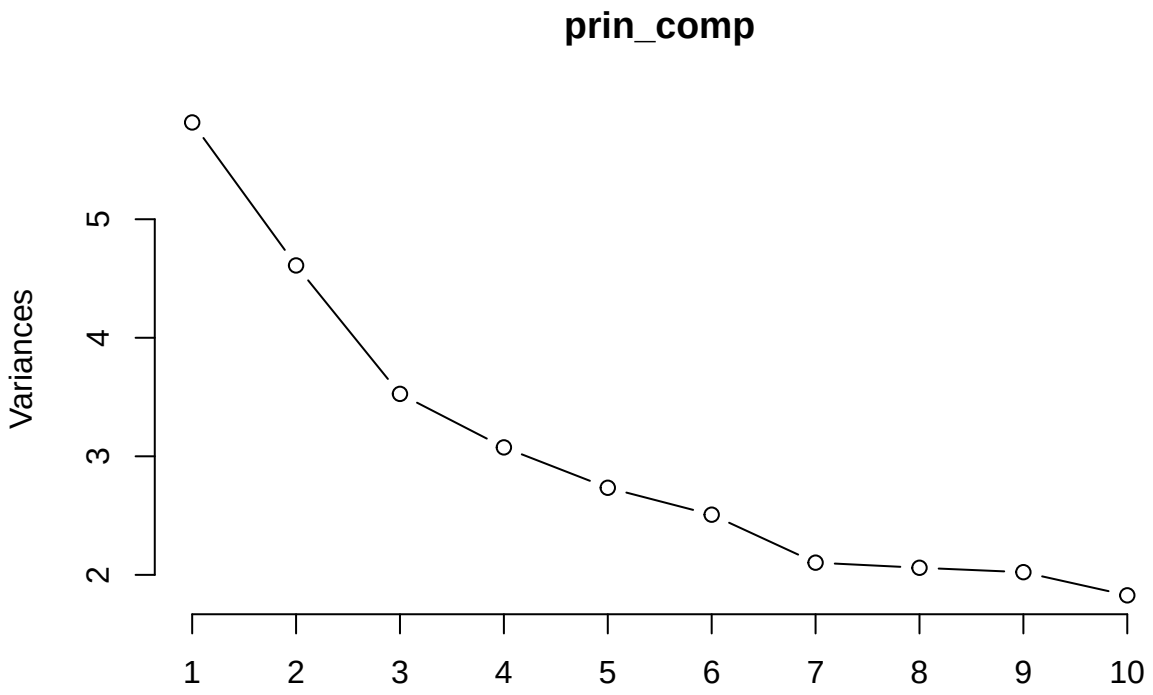
2019-1-30

How about clustering the PCs instead of variables?

Apply PCA on the dataset.

Draw variance plot:

```
plot(prin_comp,type = 'l')
```



Maybe we could select first 4 pcs?

```
pca_data = prin_comp$x[,1:4]
pca_data = as.data.frame(pca_data)
pca_data$ProjectSpecificId = merged$ProjectSpecificId
pca_data$trt = merged$trt
```

```
dim(pca_data)
```

```
## [1] 150   6
```

```
head(pca_data)
```

```
##          PC1        PC2        PC3         PC4 ProjectSpecificId trt
## 1 -0.5717660 -1.6985164  0.2976318  2.48851509            CU0011   1
## 2  0.4101561 -2.7930809 -1.0161340  0.98530249            CU0014   1
## 3  3.3495284  3.0639121  1.1848662  0.14257675            CU0016   2
## 4 -1.6587150  0.8456167 -1.3515450 -0.07077259            CU0022   2
## 5  0.8730392 -0.5039017 -1.9045270 -0.16311165            CU0024   2
## 6  0.7742141 -1.9016804 -1.2662060  3.38649196            CU0027   2
```

Apply k-means to cluster the pcs

k = 2

```
km_pca = kmeans(pca_data[,1:4], 2, nstart = 25)
```

The VI value:

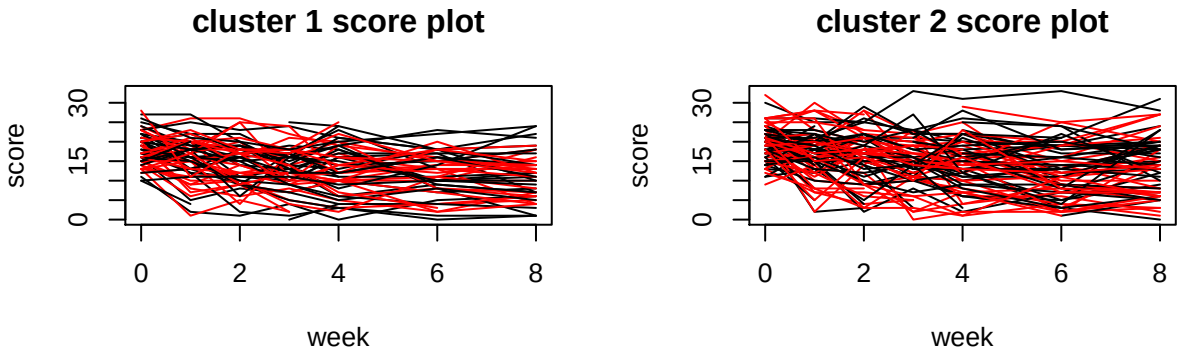
```
vi(cbind(pca_data$trt, km_pca$cluster))
```

```
## [1] 1.370982
```

```
kable(cluster_summary_table(km_pca,pca_data))
```

	cluster 1	cluster 2	Total
Drug	31	45	76
Placebo	31	43	74
Total	62	88	150

```
cluster_score_plot2(km_pca,pca_data)
```



k = 4

```
km_pca = kmeans(pca_data[,1:4], 4, nstart = 25)
```

The VI value:

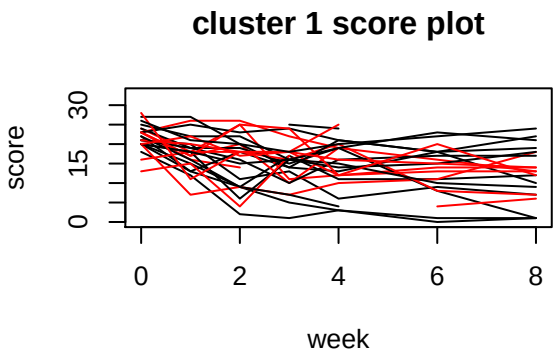
```
vi(cbind(pca_data$trt, km_pca$cluster))
```

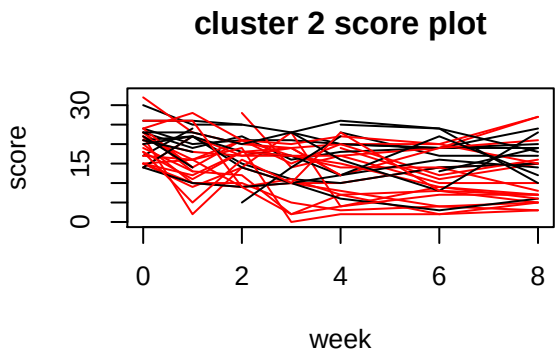
```
## [1] 2.036596
```

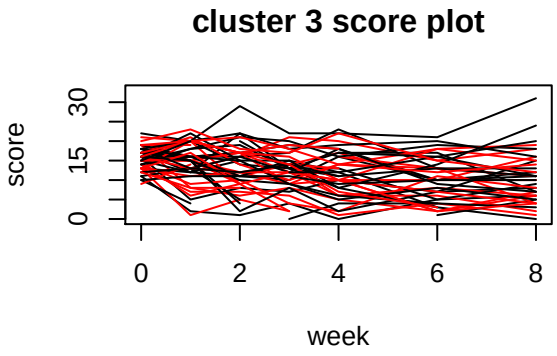
```
kable(cluster_summary_table(km_pca,pca_data))
```

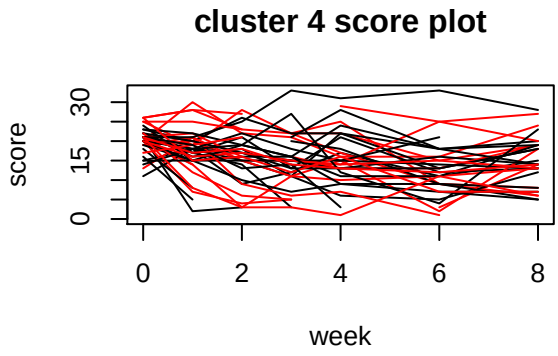
	cluster 1	cluster 2	cluster 3	cluster 4	Total
Drug	11	20	25	20	76
Placebo	15	13	25	21	74
Total	26	33	50	41	150

```
cluster_score_plot2(km_pca,pca_data)
```


cluster 1 score plot
score
0
15
30
0
2
4
6
8
week


cluster 2 score plot
score
0
15
30
0
2
4
6
8
week


cluster 3 score plot
score
0
15
30
0
2
4
6
8
week


cluster 4 score plot
score
0
15
30
0
2
4
6
8
week