Clustering Patient Trajectories

Abby Flynt

Carnegie Mellon University
Department of Statistics

Howard Seltman, Rebecca Nugent, and Joel Greenhouse

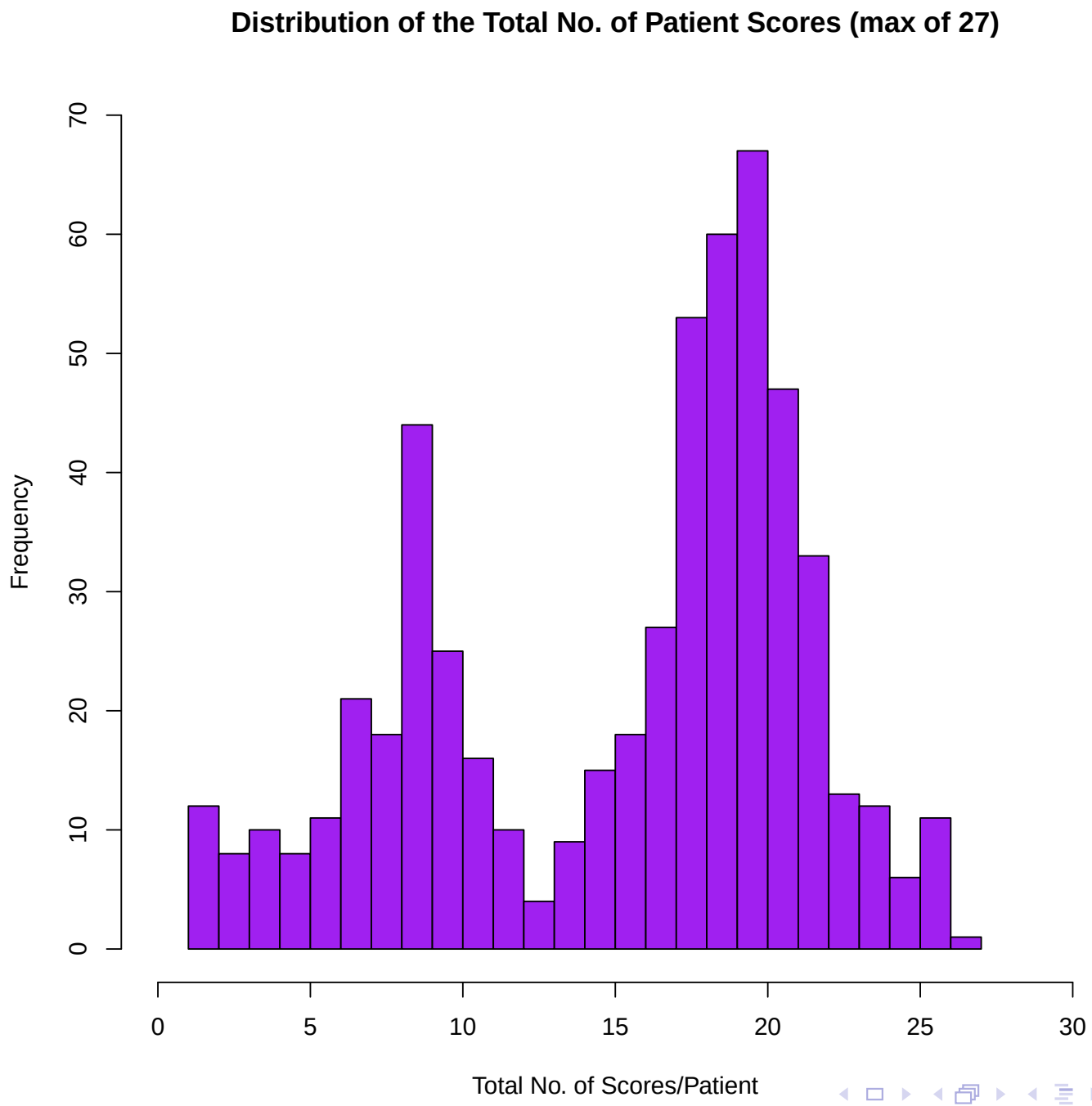
Spring 2010

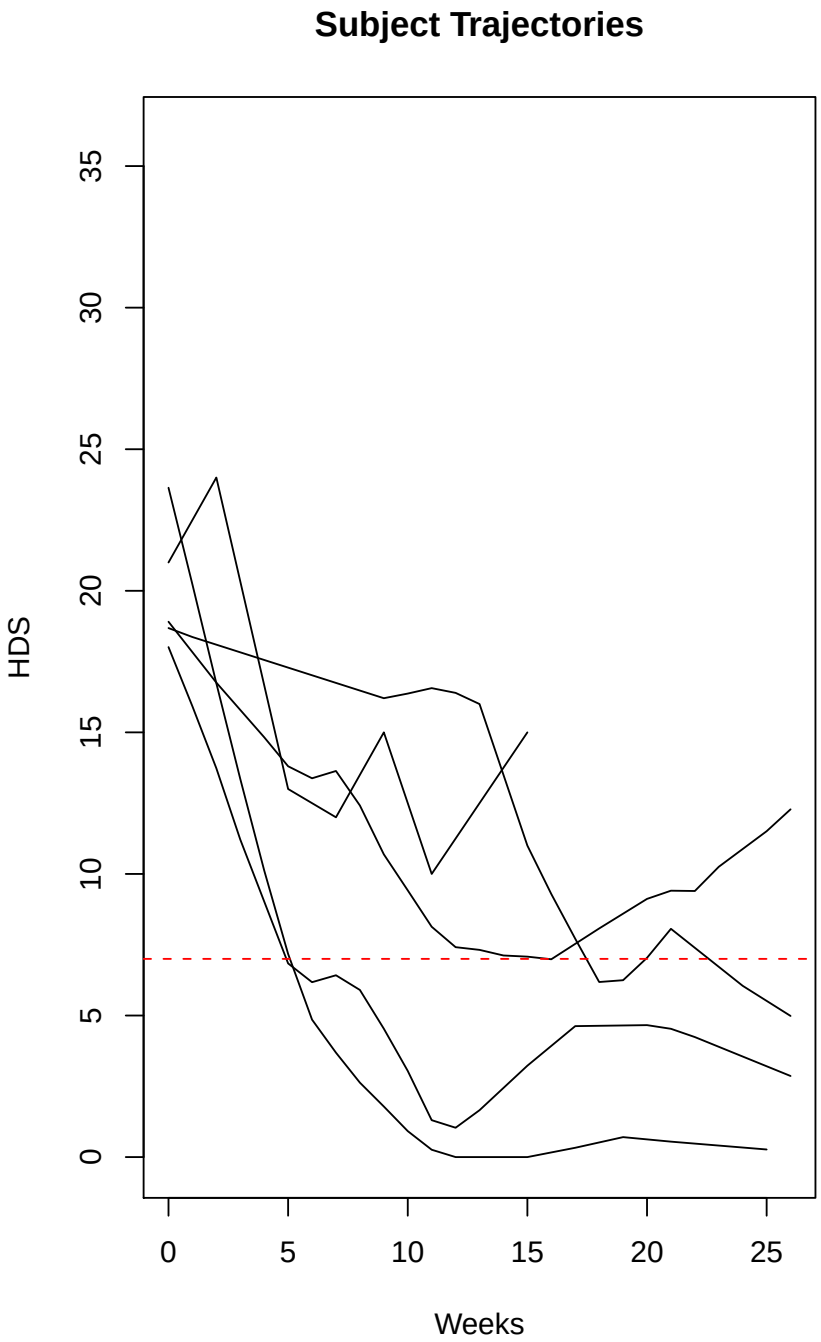
The Hamilton Rating Scale for Depression (HDS)

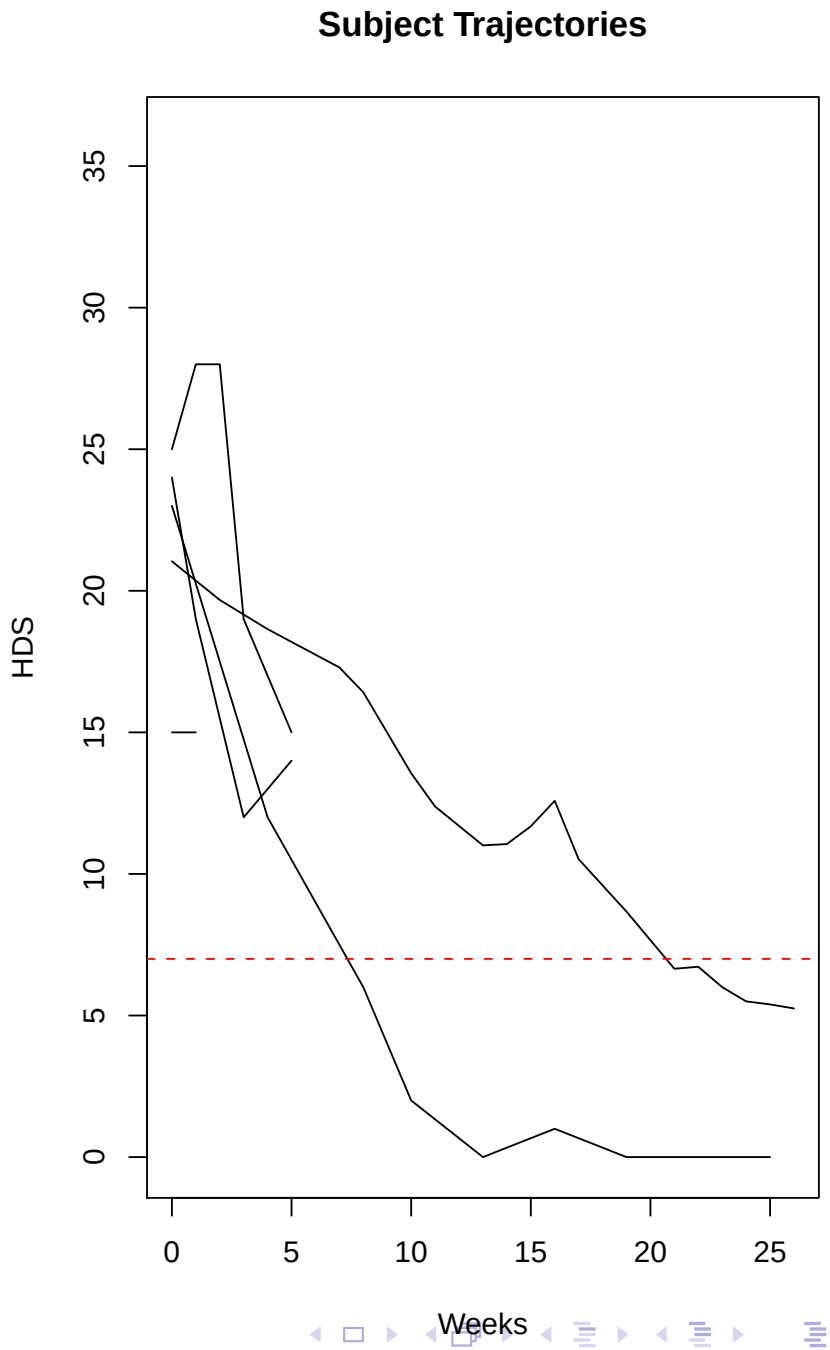
- Used since 1960 to rate the severity of a person's depression
- Answers to first 17 (out of 21) questions comprise the score
- Any score that is greater than 7 (out of a max of 53) indicates depression

The Data

- The data come from the Pittsburgh 600 Study
- Patients were in one of 5 studies:
  - Maintenance Therapies of Recurrent Depression (DPP, 230)
  - Social Zeitgebers in Depression (SZ, 93)
  - Maintenance Therapies of Late-Life Depression (MTLD, 126)
  - Psychobiology of Recovery from Depression (PRD, 56)
  - Nocturnal Penile Tumescence in Depression Study (NPT, 54)
- Have a weekly HDS for each patient from 0 weeks to 26 weeks
- Also have the patients age and gender among other things

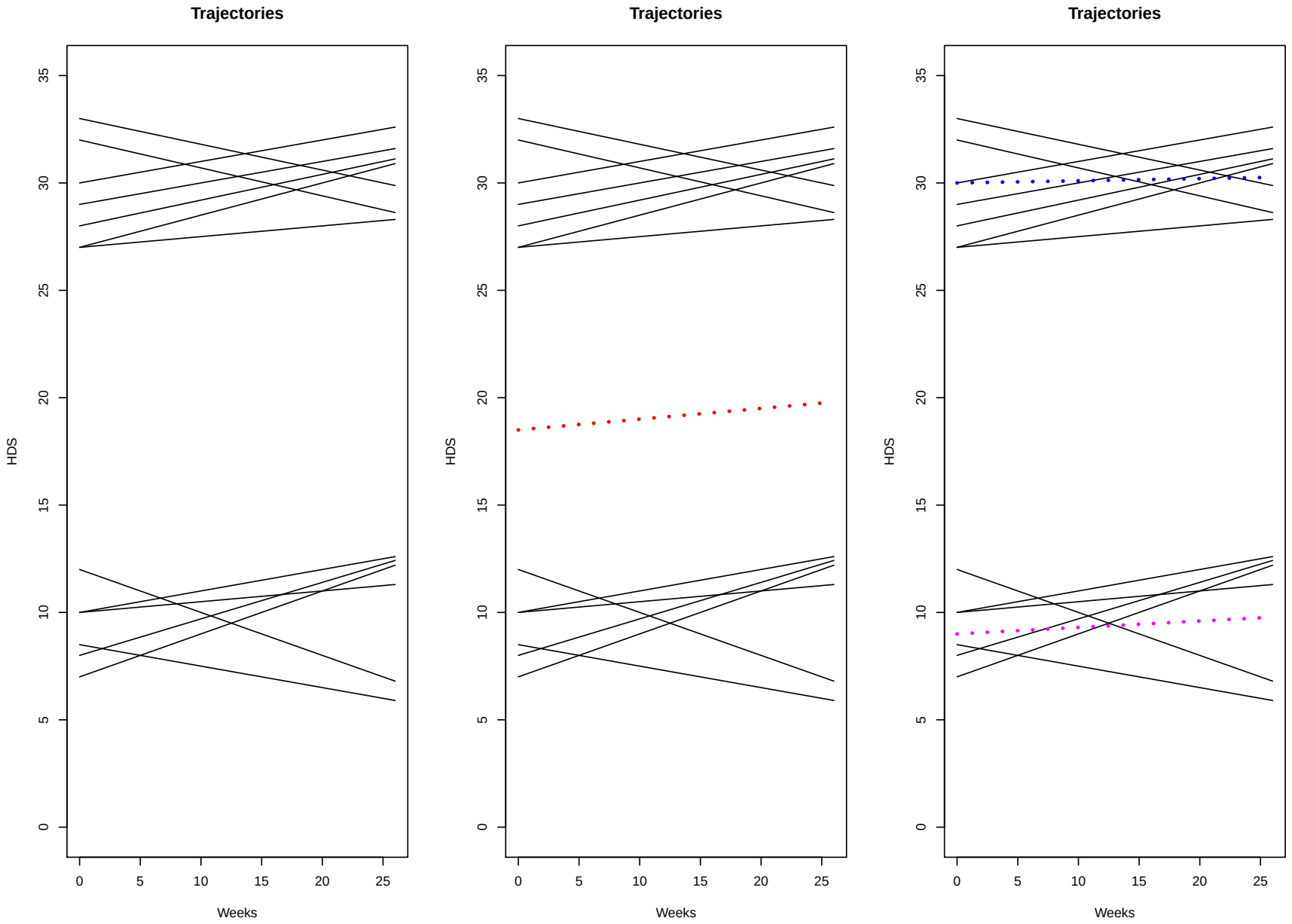

Distribution of the Total No. of Patient Scores (max of 27)
Frequency
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
Total No. of Scores/Patient


Subject Trajectories
HDS
Weeks
0
5
10
15
20
25
30
35


Subject Trajectories
HDS
Weeks
0
5
10
15
20
25
30
35

The Idea

- The general idea is to cluster together similar patient trajectories
- Ones that "look alike"
- Why? So we can predict who will "recover" based on their early scores
- A recovery is defined as having an HDS that remains at 7 or below.


Trajectories
HDS
Weeks
0
5
10
15
20
25
30
35
Trajectories
HDS
Weeks
Trajectories
HDS
Weeks

FLEXMIX

- Clustering with the R package flexmix
- Flexmix uses the EM algorithm to cluster trajectories
- Iterates between computing the expectation of the log likelihood and maximizing it to find the parameters of the clusters
- Converges when the difference in parameter values from step i to step $i + 1$ is small.

FLEXMIX

- Different ways to choose the number of clusters
  - Can give how many you want (like k-means)
  - Can give it a range of values and decide which is best (BIC)
- Also want to give it starting values for the EM

FLEXMIX MODELS

```
step.log1 = stepFlexmix(HDS~logWeeks+I(logWeeks^2)+
                    scale(Age, scale=F)+Sex|ID,
                    data=pgh.sub1,
                    model = FLXMRglm(family="gaussian"),
                    k = 1:10, nrep=20)
```

```
step.log25 = stepFlexmix(HDS~logWeeks+I(logWeeks^2)+
                    Study+scale(Age, scale=F)+Sex|ID,
                    data=pgh.sub25,
                    model = FLXMRglm(family="gaussian"),
                    k = 1:10, nrep=20)
```

Plots

Time for some fun plots!!