Homework 4
Applied Mutlivariate Analysis

Emorie Beck

September 22, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(kableExtra)
library(multcomp)
library(lme4)
library(plyr)
library(tidyverse)
library(MVN)
```

1.2 data

The file, Set_4.csv, contains data from a study in which college students completed the 10-item Rosenberg Self-Esteem Scale on two occasions spaced 4 weeks apart. The Rosenberg Scale contains the following items, rated using a scale that ranged from 1 (Strongly Agree) to 4 (Strongly Disagree):

1. I feel that I am a person of worth, at least on an equal basis with others.
2. I feel that I have a number of good qualities.
3. All in all, I am inclined to feel that I am a failure.
4. I am able to do things as well as most other people.
5. IfeelIdonothavemuchtobeproudof.
6. I take a positive attitude about myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself. 9. I certainly feel useless at times.
9. At times I think I am no good at all.

The items in the data file have been reversed where necessary so that higher numbers reflect higher self-esteem. The scale is assumed to have a single underlying dimension.

```r
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework4"

dat <- sprintf("%s/Set_4.csv", wd) %>%
  read.csv(., stringsAsFactors = F)  %>%
  mutate(SID = 1:n())

head(dat)
```

```
##   Item_1_Time_1 Item_2_Time_1 Item_3_Time_1 Item_4_Time_1 Item_5_Time_1
## 1             4             4             4             3             4
## 2             4             4             4             3             4
## 3             3             3             3             3             1
## 4             4             4             4             4             4
## 5             3             3             4             3             3
## 6             3             3             3             3             2
##   Item_6_Time_1 Item_7_Time_1 Item_8_Time_1 Item_9_Time_1 Item_10_Time_1
## 1             3             3             3             4              4
## 2             3             3             3             3              4
## 3             2             2             1             2              2
## 4             3             4             3             3              4
## 5             3             2             4             4              4
## 6             3             2             1             1              1
##   Item_1_Time_2 Item_2_Time_2 Item_3_Time_2 Item_4_Time_2 Item_5_Time_2
## 1             4             4             4             3             4
## 2             4             4             4             3             4
## 3             3             3             4             3             3
## 4             3             4             3             3             4
## 5             3             3             3             3             2
## 6             3             3             3             3             4
##   Item_6_Time_2 Item_7_Time_2 Item_8_Time_2 Item_9_Time_2 Item_10_Time_2
## 1             4             3             4             4              4
## 2             3             3             4             3              4
## 3             2             2             1             2              3
## 4             3             3             2             3              4
## 5             3             3             4             4              4
## 6             2             2             2             1              1
##   SID
## 1   1
## 2   2
## 3   3
## 4   4
## 5   5
## 6   6
```

Answer the following questions about these data:

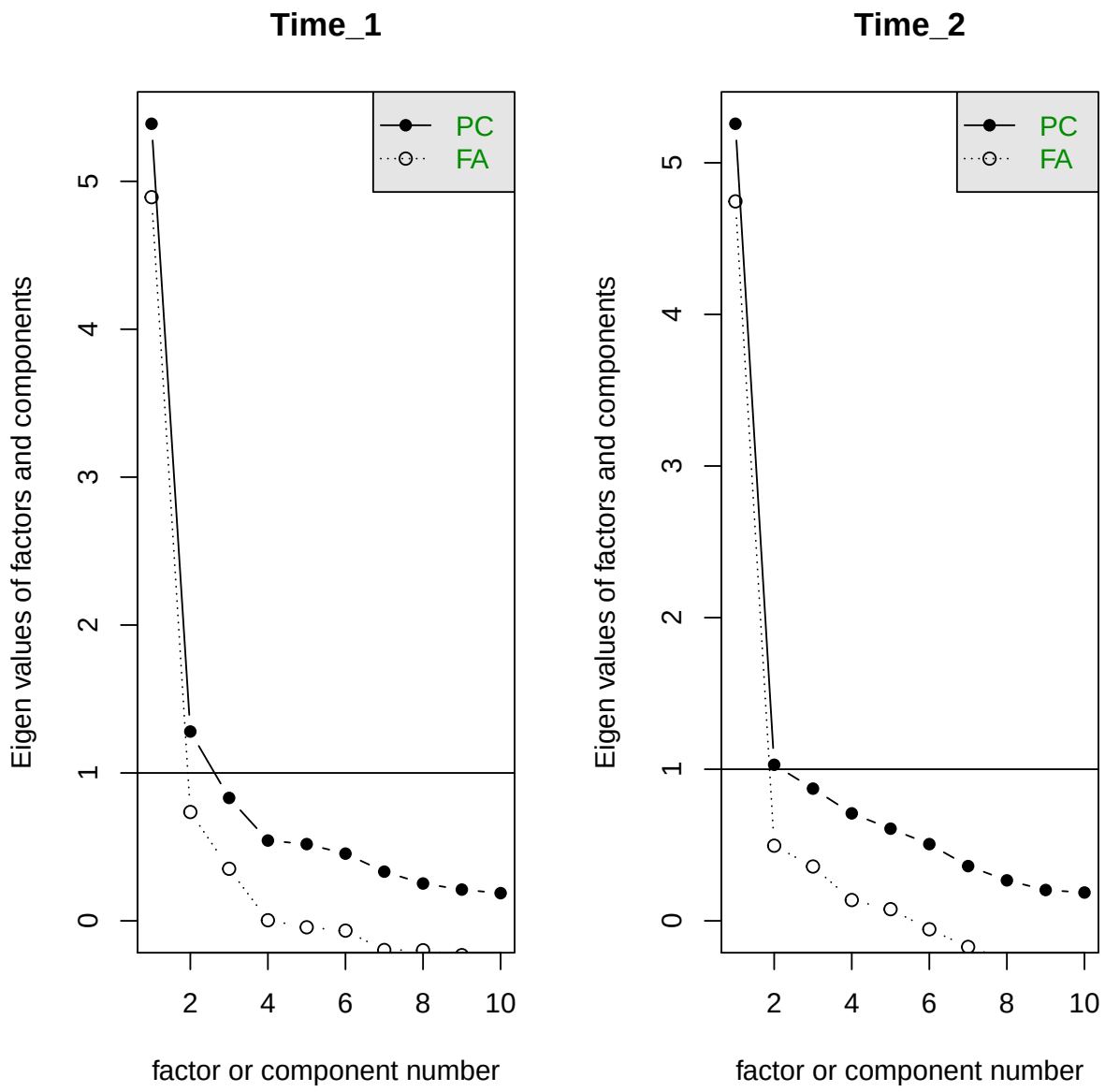
2 Question 1

For each set of 10 items, use the scree test, parallel analysis, Very Simple Structure (VSS), and Minimum Average Partial (MAP) to determine the appropriate number of principal components to extract. Note that these different ways of determining the appropriate number of components need not agree. In your opinion, is the unidimensionality assumption supported by these tests?

2.1 Scree Test

```
nested_dat <- dat %>%
  gather(key = item, value = value, -SID) %>%
  separate(item, c("Item", "ItemNum", "Time", "TimeNum"), sep = "_") %>%
  unite(item, Item, ItemNum, sep = "_") %>%
  unite(time, Time, TimeNum, sep = "_") %>%
  spread(item, value) %>%
  select(-SID) %>%
  group_by(time) %>%
  nest()

par(mfrow = c(1,2))
nested_dat %>%
  mutate(scree = map2(data, time, ~scree(.x, main = .y)))
```

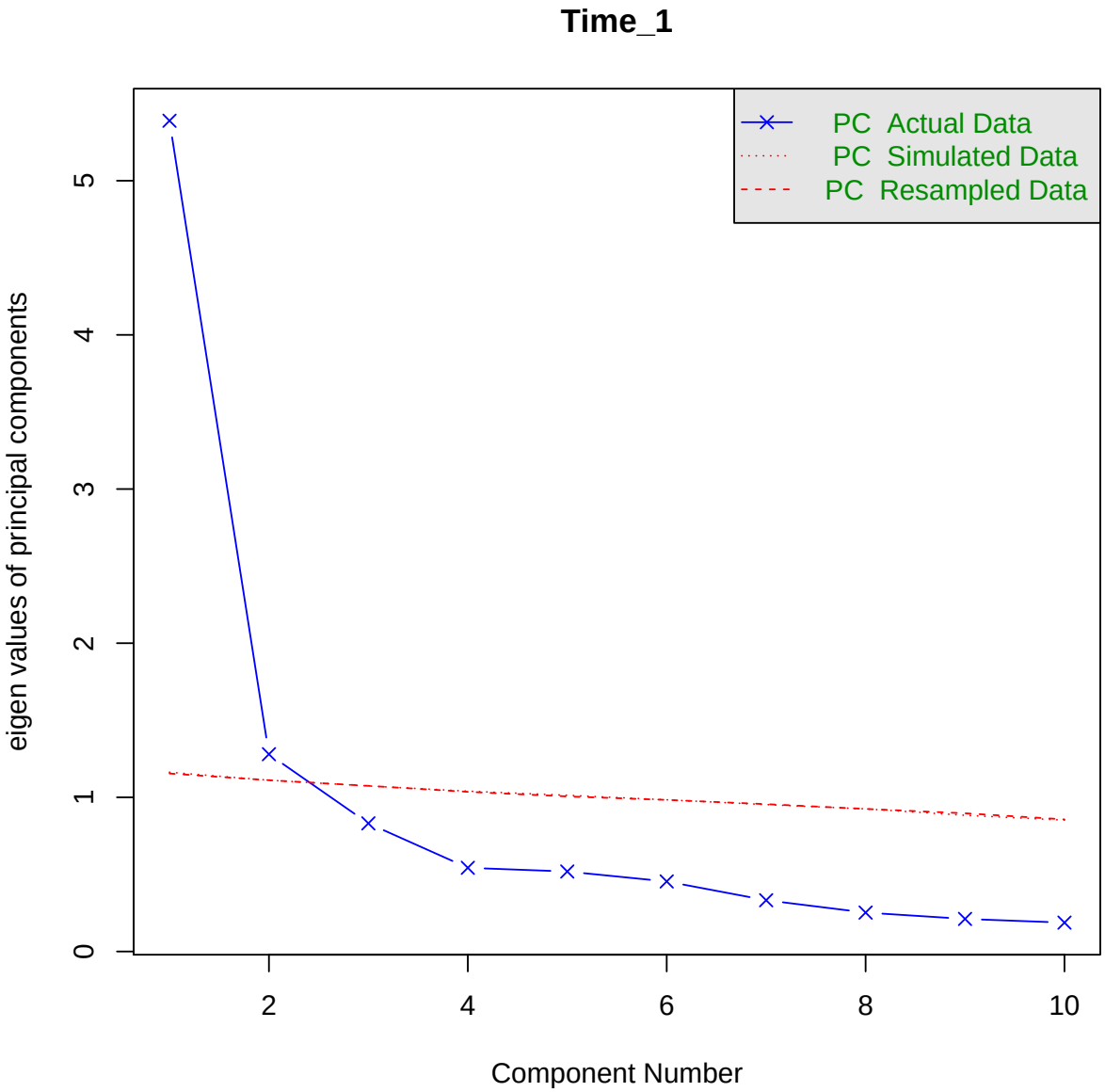



```
## # A tibble: 2 x 3
##   time   data                 scree
##   <chr>  <list>               <list>
## 1 Time_1 <tibble [1,000 x 10]> <S3: psych>
## 2 Time_2 <tibble [1,000 x 10]> <S3: psych>
```

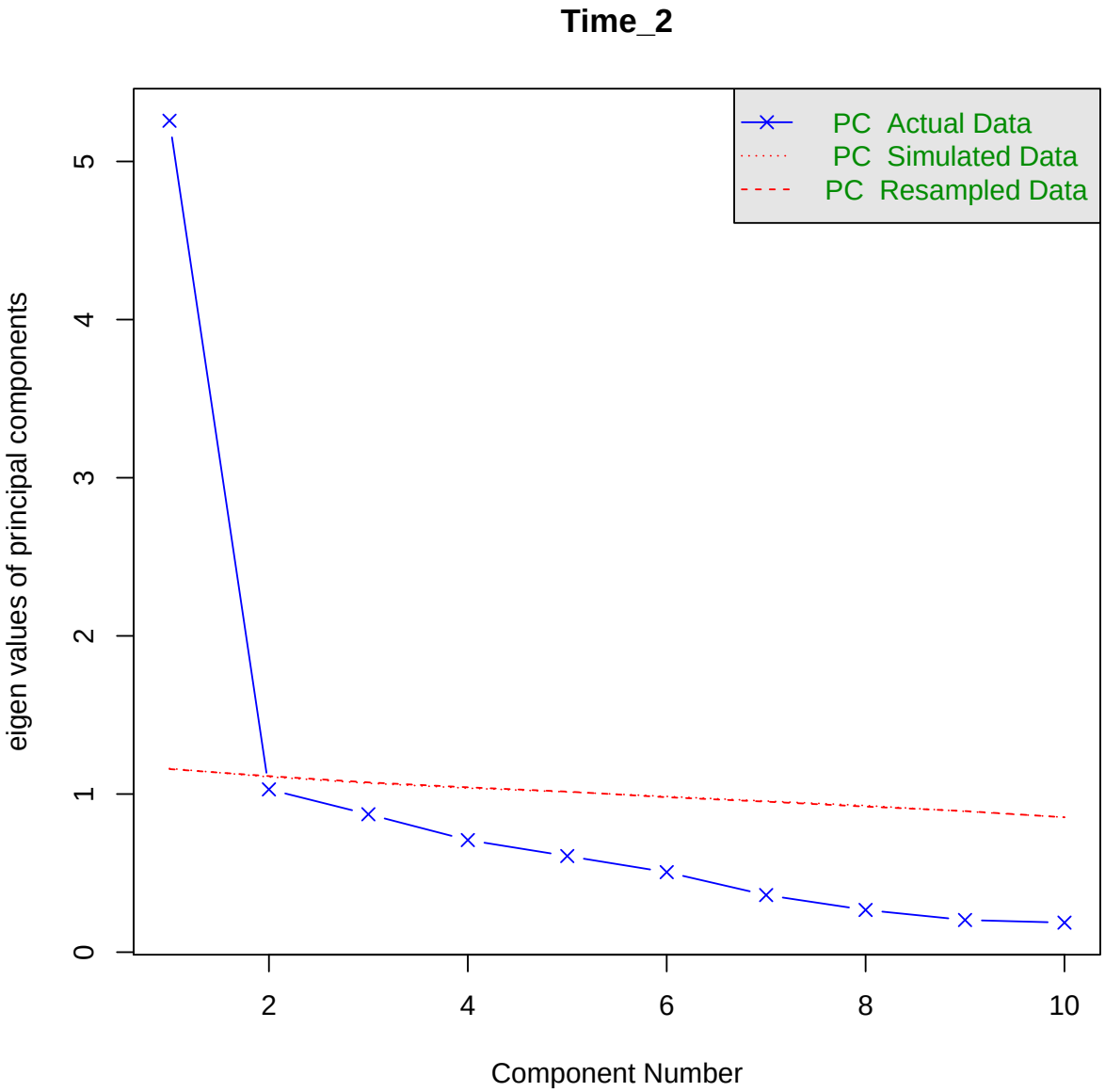
The scree test suggests 2 components at time 1 and 2.

2.2 Parallel Analysis

```
nested_dat %>%
  mutate(parallel = map2(data, time, ~fa.parallel(.x, main = .y, fa = "pc")))
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  2
```

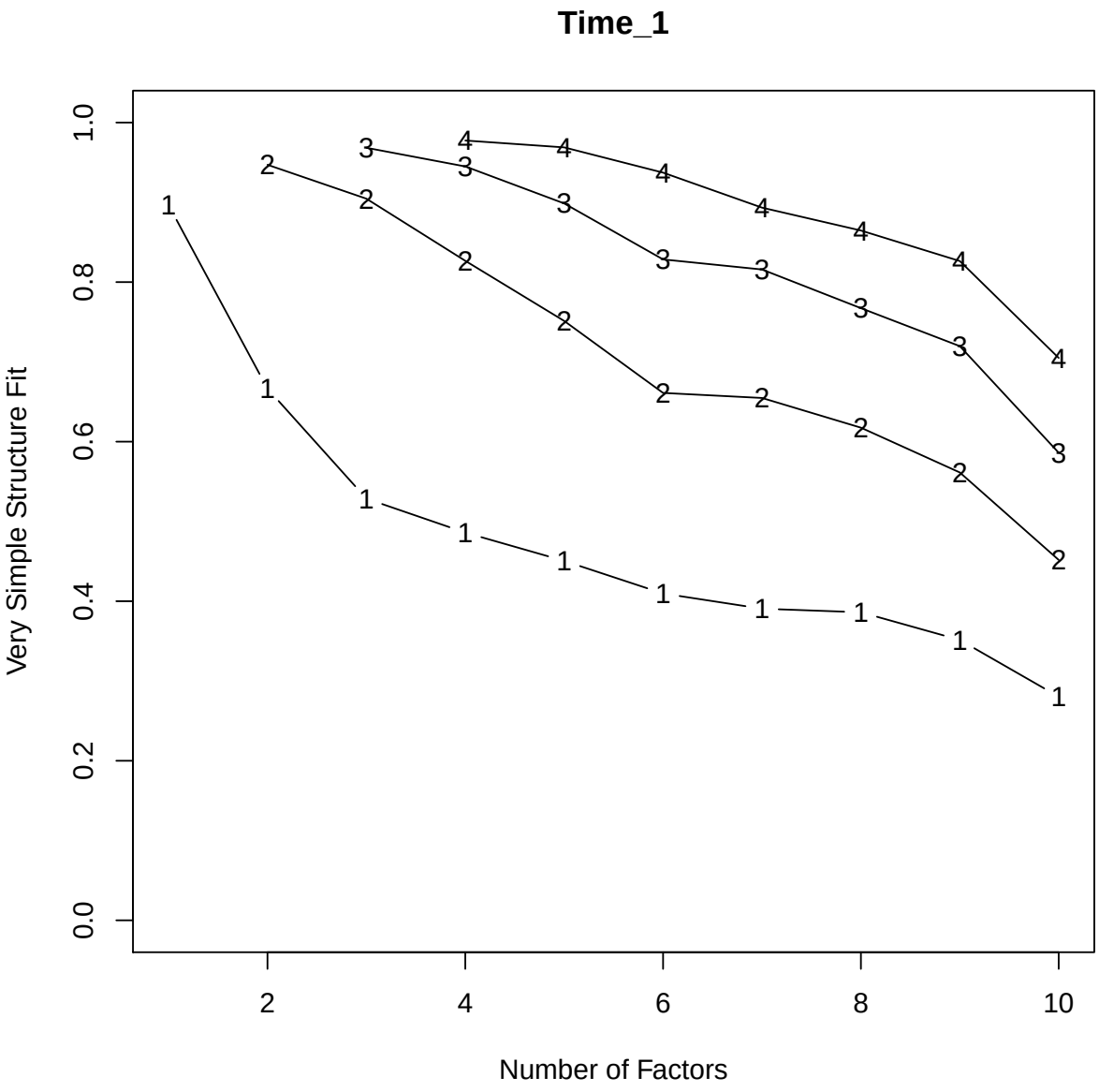


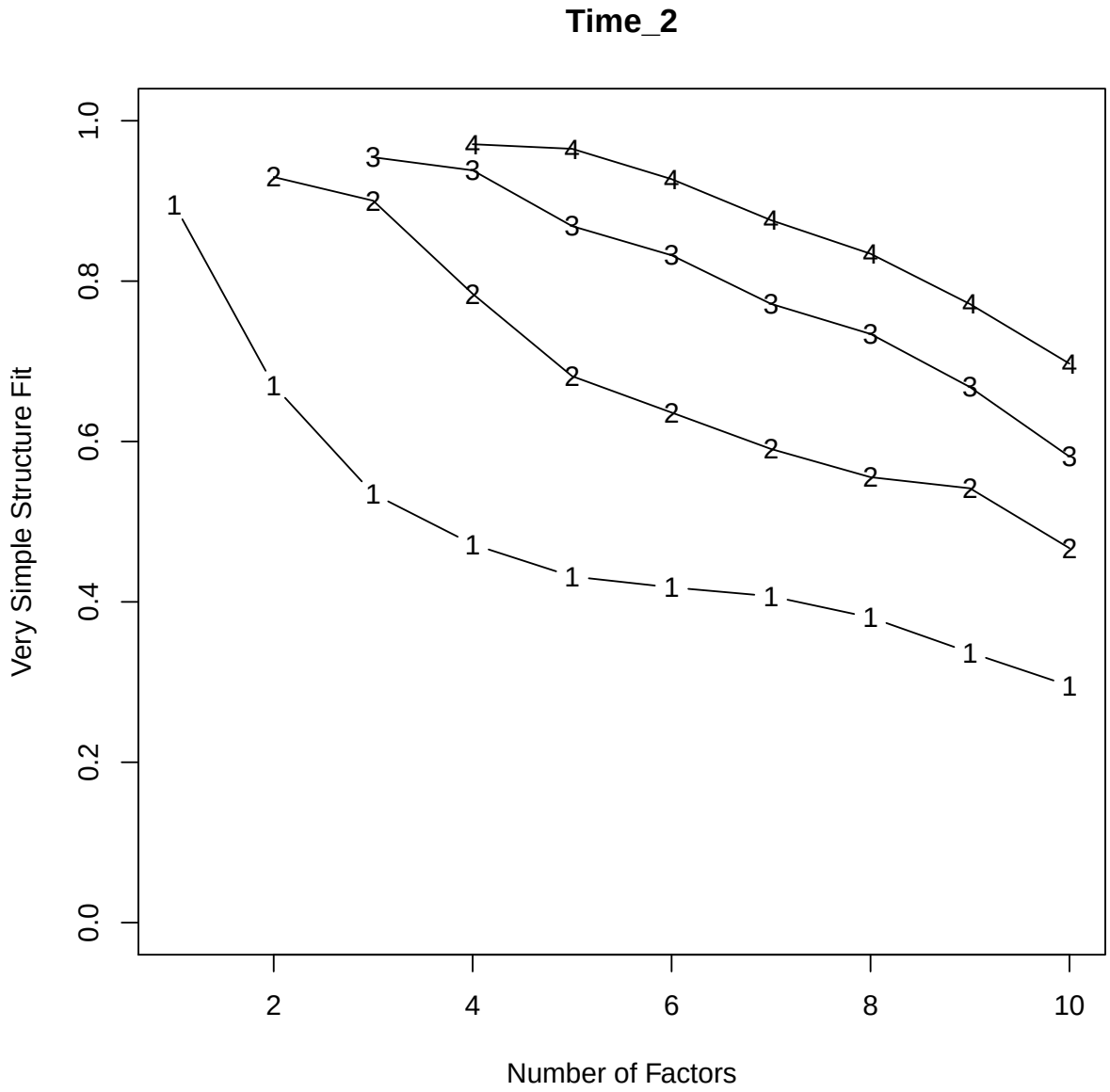

```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  1
## # A tibble: 2 x 3
##   time   data                 parallel
##   <chr>  <list>               <list>
## 1 Time_1 <tibble [1,000 x 10]> <S3: psych>
## 2 Time_2 <tibble [1,000 x 10]> <S3: psych>
```

Paralell analysis suggests 2 components at time 1 and 1 at time 2.

2.3 Very Simple Structure (VSS)

```
par(mfrow=c(1,1))
nested_dat <- nested_dat %>%
  mutate(vss = map2(data, time, ~vss(.x, n = 10, rotate = "none", title = .y, fm = "pc")))
```



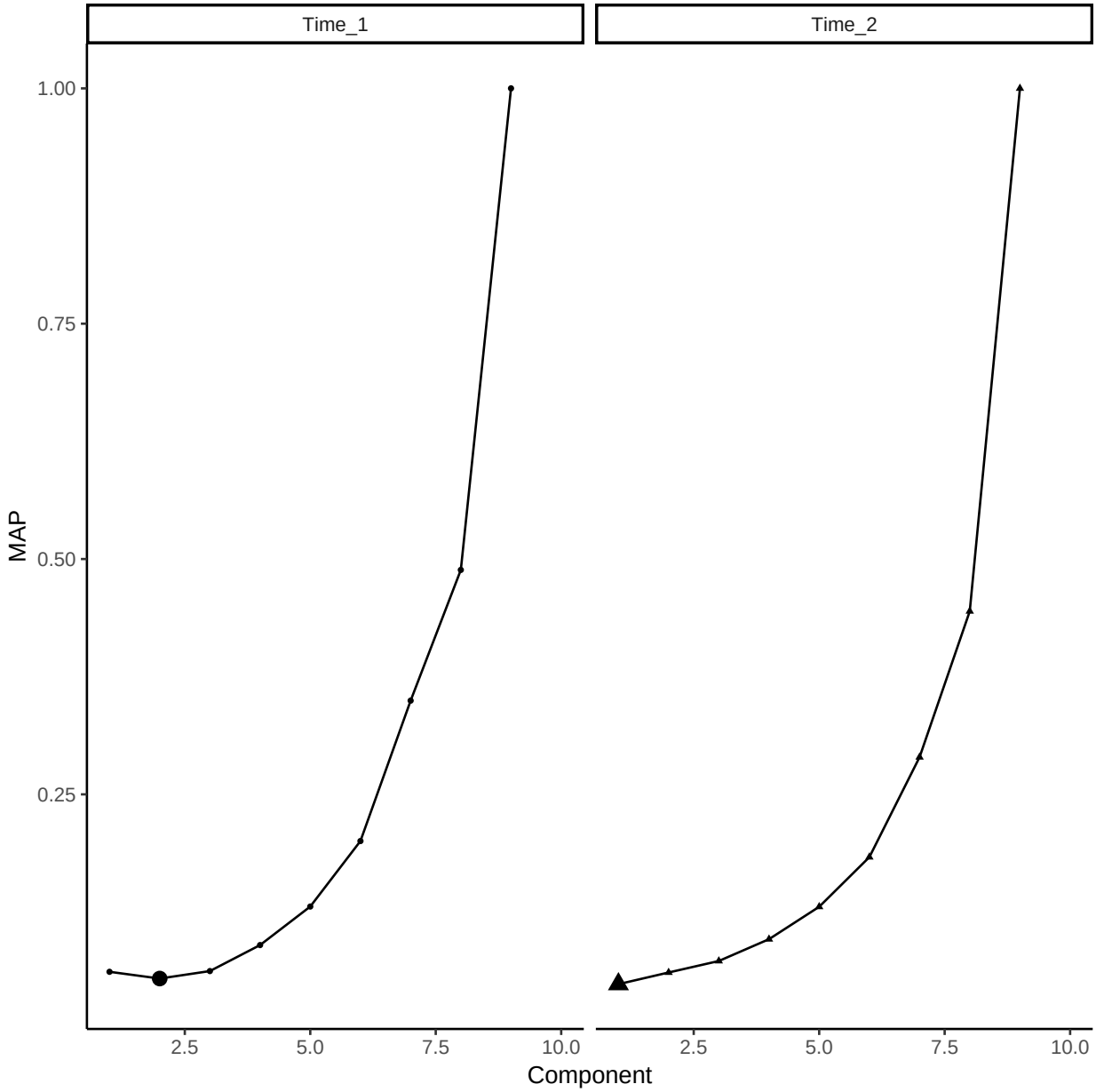


VSS seems to suggest 3 factors are optimal at both times.

2.4 Minimum Average Partial (MAP)

```
map_fun <- function(v){
  data.frame(v$map) %>% mutate(Component = 1:nrow(.)) %>%
    setNames(c("MAP", "Component"))
}
nested_dat %>%
  mutate(map = map(vss, map_fun)) %>%
  unnest(map) %>%
  group_by(time) %>%
  mutate(min = ifelse(MAP == min(MAP, na.rm = T), "NC", "No")) %>%
  ggplot(aes(x = Component, y = MAP, shape = time)) +
```

```
  scale_size_manual(values = c(3, 1)) +
  geom_line() +
  geom_point(aes(size = min)) +
  facet_grid(~time) +
  theme_classic() +
  theme(legend.position = "none")
```



The MAP test suggests 2 components at time 1 and 1 at time 2.
Overall, the tests seem to suggest 2 components at time 1 and 1 at time 2.

3 Question 2

Now conduct a principal components analysis on each set of 10 items, extract three principal components from each set, and save the unrotated principal component scores. Calculate the correlations among the

combined set of principal component scores (this will be a 6 x 6 matrix).

```
nested_dat <- nested_dat %>%
  mutate(pc = map(data, ~principal(., nfactors = 3, rotate = "none", scores = TRUE)),
         scores = map(pc, ~as.tibble(.$scores)))

tmp2 <- nested_dat %>% unnest(scores) %>%
  group_by(time) %>%
  mutate(SID = 1:n()) %>%
  gather(key = item, value = value, -time, -SID) %>%
  unite(tmp, time, item, sep = ".") %>%
  spread(tmp, value)
```

3.1 Part A

What are the correlations for corresponding component scores (Time 1 and Time 2)?

```
cor(tmp2 %>% select(Time_1.PC1:Time_1.PC3), tmp2 %>% select(Time_2.PC1:Time_2.PC3)) %>%
  kable(., "latex", row.names = F, digits = 2,
        col.names = c("PC1", "PC2", "PC3")) %>%
  add_header_above(c("Time 2" = 3)) %>%
  group_rows("Time 1", 1, 3)
```

Time 2		
PC1	PC2	PC3
Time 1		
0.84	0.00	0.04
0.15	-0.38	-0.09
0.04	0.10	-0.22

3.2 Part B

The magnitudes of these correlations will be related to the magnitudes of the component eigenvalues. Why does this make sense?

Eigenvalues represent the variances of the components, so the scores should capture much of the variance of the components.

4 Question 3

Repeat the analyses, but now rotate to simple structure in the two sets using varimax rotation. Save the rotated principal component scores and examine the intercorrelations.

```
nested_dat <- nested_dat %>%
  mutate(pc.vm = map(data, ~principal(., nfactors = 3, rotate = "varimax", scores = TRUE)),
         scores.vm = map(pc.vm, ~as.tibble(.$scores)))

tmp3 <- nested_dat %>% unnest(scores.vm) %>%
  group_by(time) %>%
  mutate(SID = 1:n()) %>%
  gather(key = item, value = value, -time, -SID) %>%
  unite(tmp, time, item, sep = ".") %>%
  spread(tmp, value)
```

4.1 Part A

Do the principal components from Time 1 replicate at Time 2?

```
cor(tmp3 %>% select(Time_1.RC1:Time_1.RC3), tmp3 %>% select(Time_2.RC1:Time_2.RC3)) %>%
  kable(., "latex", row.names = F, digits = 2,
        col.names = c("PC1", "PC2", "PC3")) %>%
  add_header_above(c("Time 2" = 3)) %>%
  group_rows("Time 1", 1, 3)
```

Time 2		
PC1	PC2	PC3
Time 1		
0.11	0.30	0.35
0.58	0.02	0.32
0.32	0.41	0.10

No, the highest correlation between any 2 given factors is .66. Indeed, the third component at the second time point does not strongly resemble any component at time 1.

4.2 Part B

Why?

The correlation between factor scores at the different times is weak. The second component at time 1 appears to resemble the first component at time 2 (r = .66), but it's not clear that the 1st or 3rd components in wave are reflected in any of the components at wave 2. This likely occurs because the 1st and 3rd factors are not "real." It's also possible although unlikely that something happened over the four week period that changed the structure of self-esteem among the students.

5 Question 4

Repeat the analysis from Question 3, but now use factor analysis (set the factor method option to maximum likelihood, fm="ml"; you might also need to increase the number of iterations).

```
nested_dat <- nested_dat %>%
  mutate(fa = map(data, ~fa(., nfactors = 3, rotate = "varimax", scores = TRUE, fm = "ml")),
         scores.fa = map(pc.vm, ~as.tibble(.$scores)))
```

5.1 Part A

(a) How does this affect the eigenvalues for the extracted linear combinations?

```
tibble(
  fa_1 = nested_dat$fa[[1]]$values,
  pca_1 = nested_dat$pc.vm[[1]]$values,
  fa_2 = nested_dat$fa[[2]]$values,
  pca_2 = nested_dat$pc.vm[[2]]$values)
```

```
## # A tibble: 10 x 4
##          fa_1 pca_1    fa_2 pca_2
##         <dbl> <dbl>   <dbl> <dbl>
##  1  5.06      5.39   4.91   5.26
##  2  0.988     1.28   0.743  1.03
##  3  0.521     0.831  0.600  0.872
```

```
##  4  0.100     0.543  0.194  0.709
##  5  0.0762    0.519  0.145  0.608
##  6  0.0249    0.454  0.0542 0.506
##  7 -0.0000531 0.333 -0.0114 0.361
##  8 -0.00198   0.252 -0.0675 0.267
##  9 -0.0371    0.211 -0.0996 0.203
## 10 -0.172     0.187 -0.215  0.187

nested_dat$fa[[2]]$values

##  [1]  4.91296929  0.74344285  0.60031952  0.19421193  0.14461595
##  [6]  0.05419439 -0.01144530 -0.06754470 -0.09958244 -0.21451904
```

Some of the eigenvalues are now negative and are overall smaller in magnitude.

5.2 Part B

What happens to the pattern of correlations among factor scores?

```
tmp4 <- nested_dat %>% unnest(scores.fa) %>%
  group_by(time) %>%
  mutate(SID = 1:n()) %>%
  gather(key = item, value = value, -time, -SID) %>%
  unite(tmp, time, item, sep = ".") %>%
  spread(tmp, value)

cor(tmp4 %>% select(Time_1.RC1:Time_1.RC3), tmp4 %>% select(Time_2.RC1:Time_2.RC3)) %>%
  kable(., "latex", row.names = F, digits = 2,
        col.names = c("PC1", "PC2", "PC3")) %>%
  add_header_above(c("Time 2" = 3)) %>%
  group_rows("Time 1", 1, 3)
```

Time 2		
PC1	PC2	PC3
Time 1		
0.11	0.30	0.35
0.58	0.02	0.32
0.32	0.41	0.10

They appear unaffected. At best, only one factor appears to replicate.

6 Question 5

What does this series of analyses tell you about the stability of principal components and factor scores, the hazards of overfactoring, and the importance of replication before trusting the meaning and interpretation of scores?

This analysis suggests that principle components and factor scores may, in some circumstances, be dubiously related over time. Moreover, the more factors that are extracted, the less likely they are to replicate. This likely occurs becuase although additional factors explain more of the total variance, the propotion of variance they explain are much smaller than previously extracted factors. In this case, at best one factor appears to be stable at all, despite the fact that the various methods for testing the needed number of factors suggested 2 factors were needed.