TutorialFactor Analysis

Tutorial ini adalah bagian dari mata kuliah Fundamen Sains Data, Informatika, UII. Pada tutorial ini, kita akan mencoba mengaplikasikan model Factor Analysis (FA) pada beberapa data set.

Contoh 1 Pada contoh pertama, kita akan menggunakan sebuah data set yang berisi 300 respons siswa-siswa terhadap sebuah kuisioner berisi 6 item, tentang mata pelajaran favorit. Keenam item tersebut merepresensikan biology (BIO), geology (GEO), chemistry (CHEM), algebra (ALG), calculus (CALC), and statistics (STAT). Skala nilai setiap item adalah 1 hingga 5, bermakna dari sangat tidak suka hingga ke sangat suka.

Di sini, misalkan kita beranggapan ada dua factor yang men-*generate* data set dengan 6 item (variabel) ini. Untuk itu kita akan memodelkan sebuah FA dengan 2 factor. Yang pertama, mari kita lihat isi data set ini. Silakan sesuaikan path pada baris ke-2 menuju folder di mana rekan-rekan taruh data set `dataset_EFA.csv`.

```
library(psych)
myData <- read.csv("dataset_EFA.csv")
head(myData)
```

```
##   BIO GEO CHEM ALG CALC STAT
## 1   1   1    1   1    1    1
## 2   4   4    3   4    4    4
## 3   2   1    3   4    1    1
## 4   2   3    2   4    4    3
## 5   3   1    2   2    3    4
## 6   1   1    1   4    4    4
```

Kita dapat lihat nilai dari 6 respons pertama. Selanjutnya kita akan memodelkan FA dengan 2 factor.

```
modelFA <- fa(r = myData, nfactors = 2, rotate = "varimax", fm="minres")
modelFA$loadings
```

```
##
## Loadings:
##      MR1   MR2
## BIO  0.853 0.132
## GEO  0.780 0.151
## CHEM 0.862
## ALG        0.794
## CALC       0.967
## STAT 0.168 0.511
##
##                  MR1   MR2
## SS loadings    2.115 1.872
## Proportion Var 0.352 0.312
## Cumulative Var 0.352 0.664
```

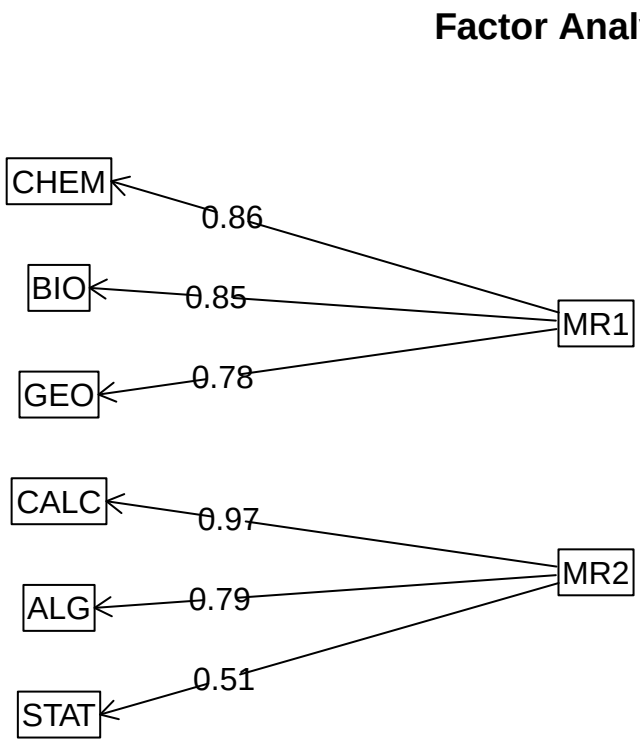
Perhatikan factor loadings dari setiap variabel ke dua factor (dengan label `MR` dan `MR2`). Secara default, `modelFA$loadings` hanya menampilkan factor loadings yang relatif besar. Untuk melihat semua factor loading lebih detailnya, dapat melalui script `modelFA$loadings[, 1:2]`.

Dapat kita lihat mata pelajaran Biology, Geology, dan Chemistry dipengaruhi oleh satu faktor yang sama (`MR1`), sedangkan Algebra, Calculus, dan Statistics dipengaruhi oleh faktor yang sama (`MR2`). Secara intuitif, kita dapat melihat jika faktor `MR1` merepresentasikan konsep *Science* dan `MR2` merepresentasikan konsep *Math*.

Dari baris `Proportion Var` dapat kita lihat besaran variance yang dijelaskan masing-masing faktor, dan jika ditotal kita mendapatkan sekitar 66% total variance. Ini dapat dimaknai bahwa model FA yang kita buat di atas membawa/merepresentasikan 66% informasi yang dikandung data set yang kita analisis. Detil output dari model FA di atas dapat dilihat melalui pemanggilan `modelFA`.

Selanjutnya, kita dapat memvisualisaikan model FA melalui script di bawah ini,

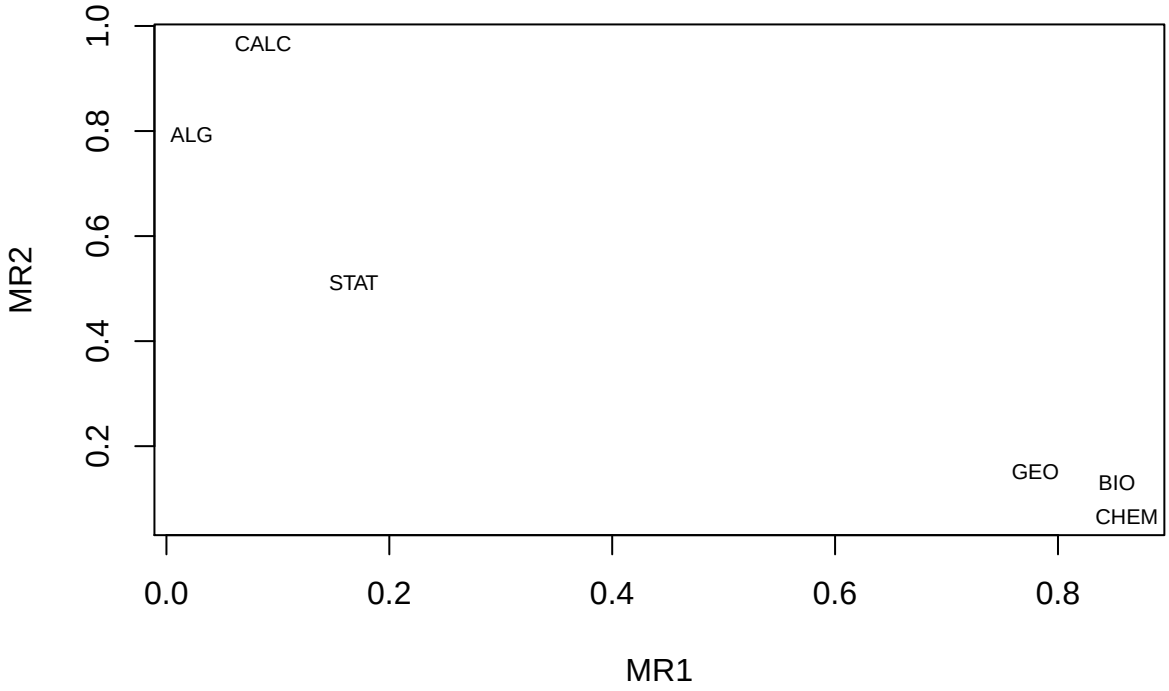
```
load <- modelFA$loadings[,1:2]
fa.diagram(modelFA, digits = 2)
```



Model FA di atas memvisualisasikan hanya factor loading yang terbesar dari setiap variabel.

Selanjutnya kita juga dapat memvisualisasikan factor loadings melalui,

```
plot(load, type="n")
text(load,labels=names(myData),cex=.7)
```



Dapat kita lihat, bagaimana beberapa mata kuliah memiliki factor loadings yang relatif tinggi pada satu faktor tertentu.

Menentukan jumlah faktor Pertanyaan ini sama ketika ketika harus menentukan jumlah k pada clustering k-means. Jika kita memiliki *preference* atau basis pengetahuan untuk menentukan jumlah faktor, maka kita dapat menggunakan pengetahuan tersebut. Jika kita tidak memiliki dasar ilmiah untuk menentukan jumlah faktor, kita dapat menggunakan **Cattell scree plot**.

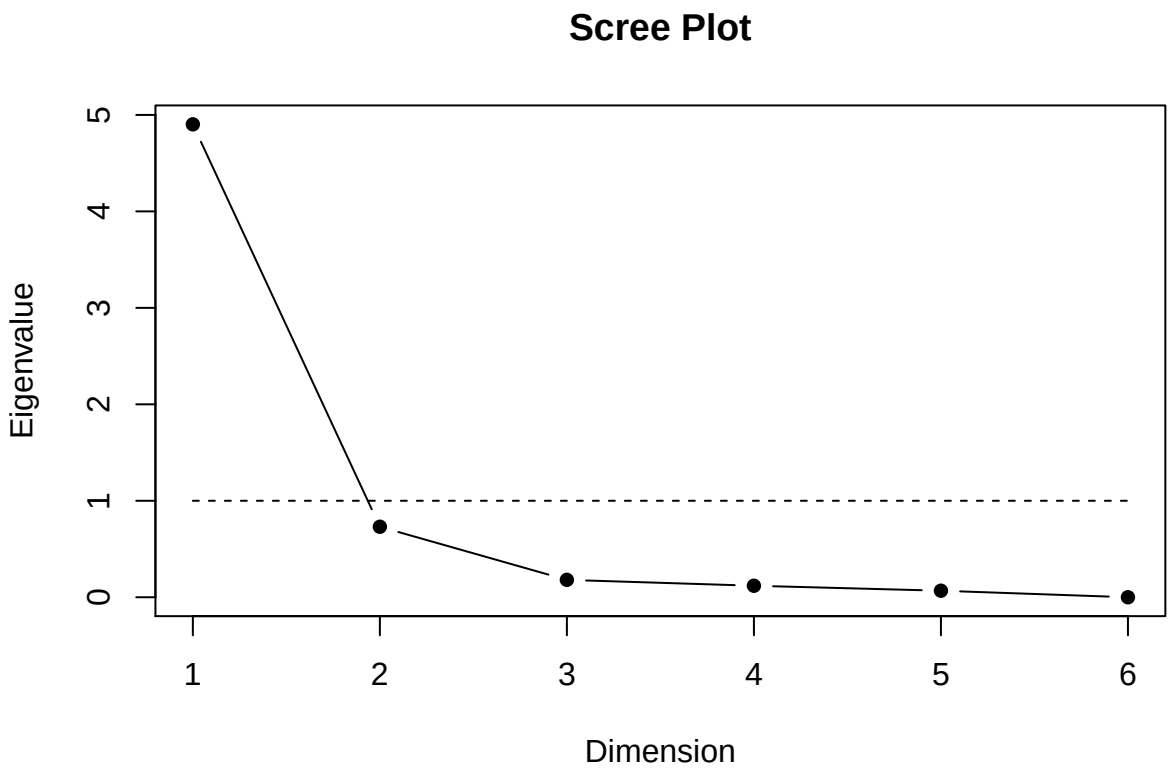
Menggunakan data set yang sama di atas, kita dapati scree plot sebagai berikut.

```
library(psy)
```

```
##
## Attaching package: 'psy'
```

```
## The following object is masked from 'package:psych':
##
##     wkappa
```

```
scree.plot(cor(myData))
```



Scree plot memvisualisasikan *eigenvalues* jika kita menggunakan satu faktor hingga enam faktor (sejumlah item/variabel di dalam data set). Bagian plot yang berbentuk *elbow* atau siku adalah batas yang dapat kita gunakan sebagai jumlah faktor. Pada scree plot di atas, bentuk siku kita temukan pada jumlah faktor = 2.

Kriteria lain yang dapat digunakan untuk menentukan jumlah faktor berdasarkan sebuah scree plot adalah aturan **Kaiser-Guttman**. Aturan ini merekomendasikan jumlah faktor yang ideal adalah sejumlah faktir yang memiliki eigenvalues lebih besar dari 1. Pada plot di atas, garis horizontal putus-putus mengindikasikan aturan tersebut. Menggunakan aturan Kaiser-Guttman, kita mendapati jumlah faktor = 1 sebagai jumlah yang ideal.

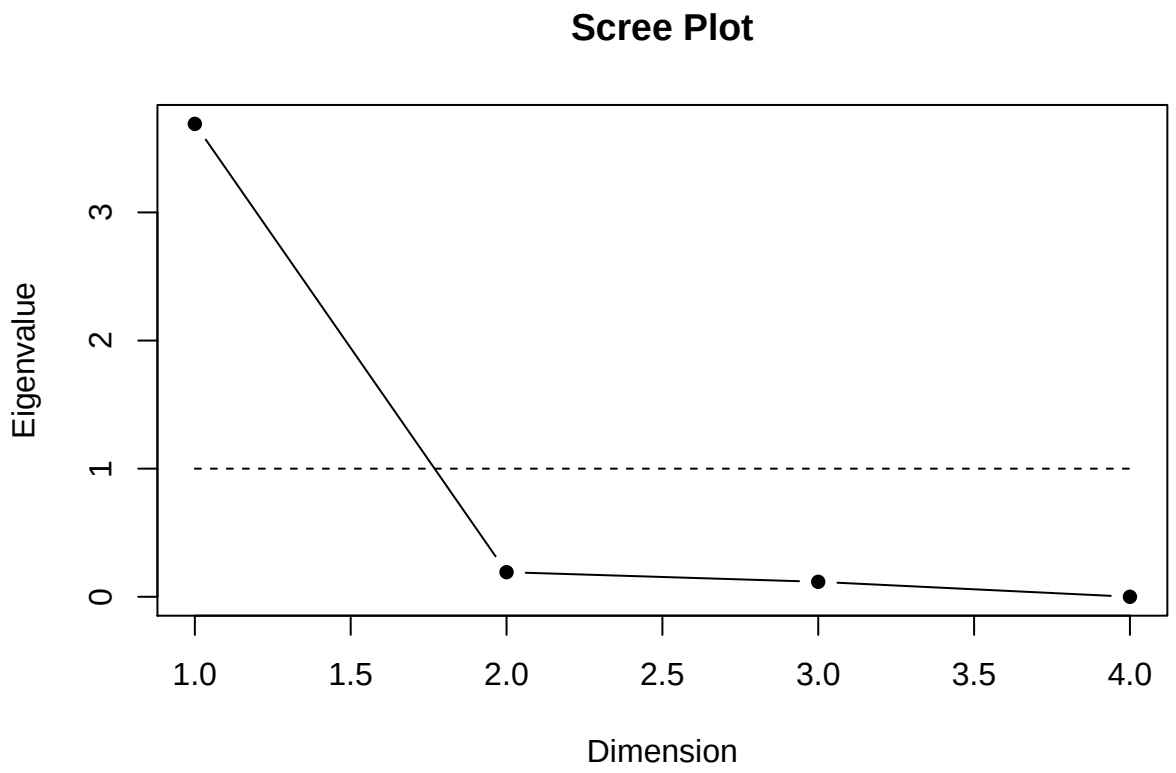
Contoh 2 Pada contoh kali ini, kita akan menggunakan sebuah data set yang mengukur suhu sebuah ruangan (dalam satuan Kelvin) pada empat sudut ruangan yang berbeda, dalam 144 waktu yang berbeda.

```
myData <- read.csv("room-temperature.csv")
head(myData)
```

```
##              Date FrontLeft FrontRight BackLeft BackRight
## 1 4/11/2010 11:30     295.2      297.0    295.8     296.3
## 2 4/11/2010 12:00     296.2      296.4    296.2     296.3
## 3 4/11/2010 12:30     297.3      297.5    296.7     297.1
## 4 4/11/2010 13:00     295.9      296.7    297.4     297.0
## 5 4/11/2010 13:30     297.2      296.5    297.6     297.4
## 6 4/11/2010 14:00     296.6      297.7    296.7     296.5
```

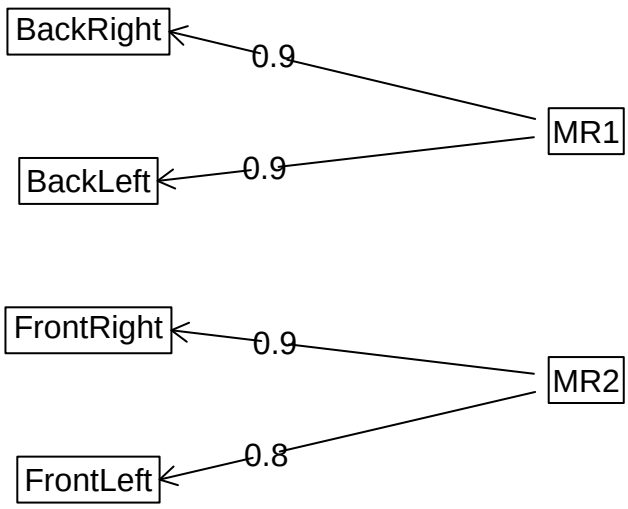
Selanjutnya kita coba gunakan scree plot untuk mendapatkan rekomendasi jumlah faktor.

```r
library(psych)
myData <- myData[, 2:5]
scree.plot(cor(myData))
```



Dari scree plot di atas, siku kita temukan pada jumlah faktor = 2, untuk itu kita akan pakai nilai ini sebagai jumlah faktor.

```r
library(psych)
faModel <- fa(myData, nfactors = 2, rotate = "varimax", fm="minres")
fa.diagram(faModel)
```



Dari model FA di atas, kita lihat suhu ruangan pada sudut belakang sebelah kanan dan kiri berasosiasi disebabkan oleh faktor yang sama, sedangkan suhu ruangan pada sudut depan kanan dan kiri disebabkan oleh faktor yang sama.

Untuk melihat detil factor loadings nya,

```
faModel$loadings
```

```
##
## Loadings:
##            MR1   MR2
## FrontLeft  0.383 0.792
## FrontRight 0.385 0.891
## BackLeft   0.866 0.286
## BackRight  0.914 0.323
##
##                  MR1   MR2
## SS loadings    1.881 1.605
## Proportion Var 0.470 0.401
## Cumulative Var 0.470 0.872
```

Total variance yang dijelaskan oleh model FA ini adalah 0.470 + 0.401 atau sekitar 87%.

Latihan

1. Download data set dari 2495 repons dari kuisioner *Generic Conspiracist Beliefs Scale* melalui link (harus dalam posisi online)

```
url <- "https://assets.datacamp.com/production/repositories/2136/datasets/869615371e66021e97829feb7e19e3
theData <- readRDS(gzcon(url(url)))
```

2. Buat scree plot berdasarkan data di atas, dan tentukan jumlah faktor yang tepat berdasarkan scree plot tersebut. Dipersilakan menggunakan konsep siku atau Kaiser-Guttman. Jumlah faktor yang Anda tentukan, bisa jadi berbeda dengan rekan yang lain; hal ini tidak menjadi masalah, yang lebih penting adalah argumentasinya.
3. Buat model FA berdasarkan jumlah faktor yang telah ditentukan di atas dan visualisasikan model FA tersebut
4. Ceritakan secara singkat factor loadings yang diestimasi pada model tersebut.
5. Berapa total variance yang dijelaskan oleh model yang ditemukan? Apakah cukup besar, atau? Jelaskan secara singkat.