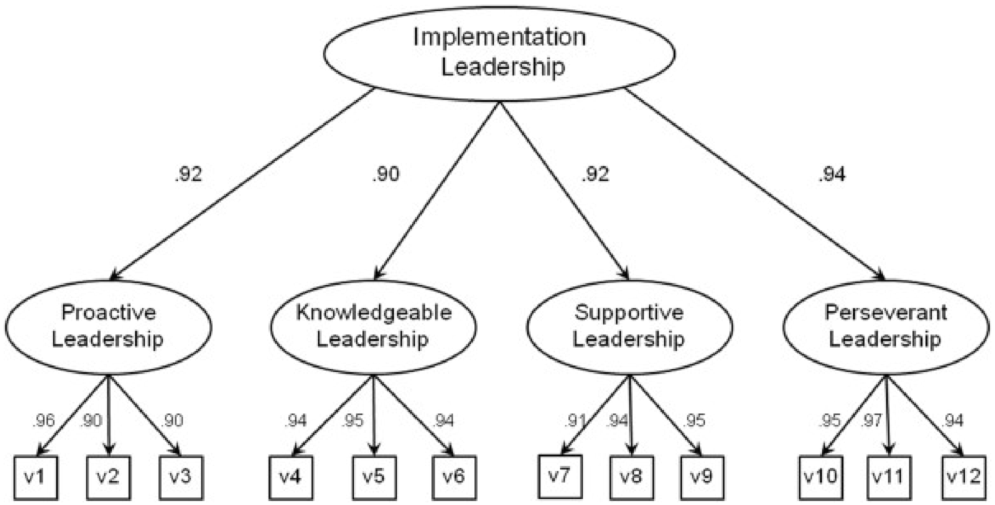
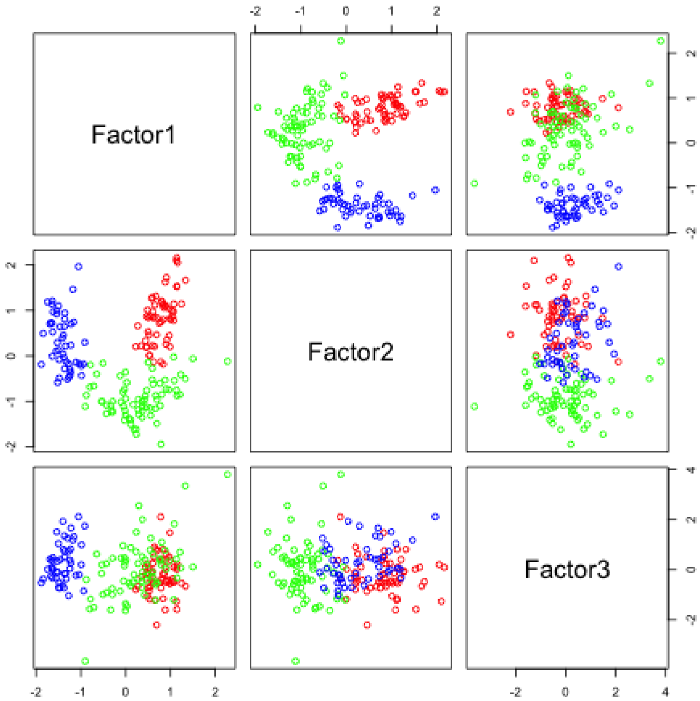
Análisis de factores





Oscar Centeno Mora

Introducción

- El análisis factorial (análisis factorial o FA) es un conjunto de técnicas cuyo orígenes se encuentran en la obra de Pearson (1901).

- Desarrollado por psicólogos donde la teoría y el método numérico no poseían justificación teórica alguna, se produjeron diversas controversias entre las escuelas de psicólogos.

- Más adelante (1940), nuevos fundamentos teóricos, numéricos y estadísticos se establecieron en las diferentes variantes del análisis factorial para ser aprobado como método de análisis.

- Algunos estadísticos que consolidaron el método: Spearman, Thomson, Thurstone, Burt, etc.

Introducción

- En el AF la atención se centra principalmente en las variables estadísticas y no en los individuos; es más un método de análisis multivariante que un método de análisis multidimensional.

- Existen dos grandes variaciones del análisis factorial: análisis factorial exploratorio, y análisis factorial confirmatorio. Nos enfocaremos únicamente en la modalidad exploratoria.

- Al igual que el PCA, en el AF se busca reducir un número p a k variables. Sin embargo, el AF posee ciertas características que lo hacen el análisis ideal para conocer las estructura subyacente que se desea estudiar. Se dice que el AF es más elaborado, y se buscan conclusiones más contundentes, el PCA busca más aproximaciones a nivel gráfico del conjunto de datos.

Diferencias entre el PCA y el AF

- De forma matemática, el PCA es un AF pero más simple - extracción de vector propio, sin rotaciones, y sin transformaciones.

- Desde un punto de vista práctico, el PCA es observacional, y el AF es un enfoque para un modelo explicito (estructura subyacente).

- Mientras que el PCA busca el análisis de la variabilidad, el AF además busca el análisis de la covariancia y correlación entre los factores.

- El objetivo típico en el análisis factorial es identificar las variables que están relacionadas entre sí, y separarlas de otras (una forma de agrupación variable). Los tipos de rotaciones ayudan en este proceso.

Etapas de un AF

- Al hacer un AF, se deberían llevar a cabo las siguientes etapas:

1. Análisis descriptivos de las variables.
2. Primera aproximación de la composición de los datos.
3. Determinar el número de las factores.
4. Segunda aproximación de la composición de los datos.
5. Ajustar diversos tipos de rotación.
6. Determinada la mejor solución en el agrupamiento de variables.

Fundamentos del AF

- El propósito del análisis factorial (AF), es describir la estructura de la matriz de variancia y covariancia entre las *p* variables de interés.

- A la reducción de variables los llamaremos ***factores***, variables subyacentes o variables latentes.

- La identificación de tales factores simplifica enormemente la descripción y comprensión de los datos multivariados. Se favorece una interpretación más sencilla y sustantiva de los datos.

- Las etapas de un ACP y AF son similares, sin embargo, en el AF buscaremos un re-trabajo en la explicación de las variables para cada uno de los factores.

Modelo del AF

- Sea $X' = (X_1, X_2, \dots, X_p)$ el vector aleatorio que contiene p variables que se pueden medir y observar, con vector de medias μ, y matriz de variancia y covariancia Σ.

- El modelo se puede presentar de forma matricial como $X - \mu = LF + \varepsilon$

donde,

L representan las cargas factoriales

F es el vector de factores comunes

ε es el vector de factores específicos

$X - \mu$ es la variable menos su media

- ¿Por qué utilizamos la forma "$X - \mu = \cdots$", y no "$X = \cdots$" ?

Modelo del AF

- Un ejemplo de un análisis factorial de un solo factor, con *p* variables es el siguiente:

$$X_1 - \mu_1 = L_1 F_1 + \varepsilon_1$$
$$X_2 - \mu_2 = L_2 F_1 + \varepsilon_2$$
$$....$$
$$X_p - \mu_p = L_p F_1 + \varepsilon_p$$

- Un ejemplo de un modelo factorial con 3 variables y 2 factores es el siguiente:

$$X_1 - \mu_1 = L_{11} F_1 + L_{12} F_2 + \varepsilon_1$$
$$X_2 - \mu_2 = L_{21} F_1 + L_{22} F_2 + \varepsilon_2$$
$$X_3 - \mu_3 = L_{31} F_1 + L_{32} F_2 + \varepsilon_2$$

Modelo del AF

- Para los datos: suponemos que hay p variables $X_1, X_2, \ldots, X_p$ que están estandarizadas.

- Se pueden estimar los F_k factores por diversas técnicas: componentes principales, máxima verosimilitud, mínimos cuadrados, etc.

- Es importante determinar el número de factores que se deben elegir.

- En el tema de PCA se escogían casi siempre 2 componentes para mejorar la representación gráfica. Acá vamos a elegir los factores según otros criterios.

Modelo del AF

- Los factores se pueden elegir según los criterios de:

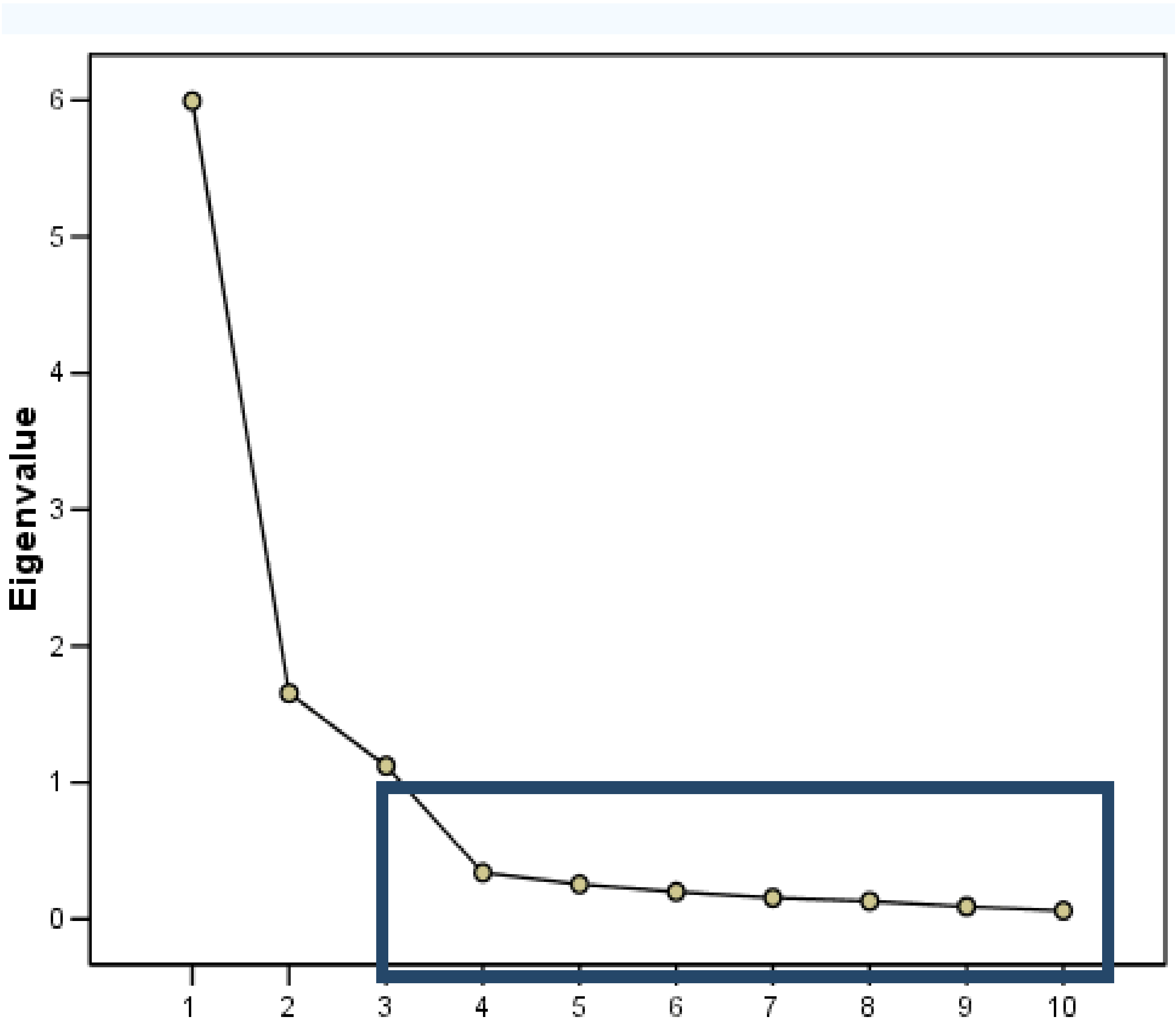
1. Elegir los k factores de modo que la matriz residual sea lo más cercana a 0. Esto es, que haya una equivalencia al encontrar k factores que haga muy pequeña la variancia explicada.

2. Retener tantos factores que sean adecuados para explicar una proporción adecuada de la variancia total.

3. Tomar los factores que tengan raíces características superiores a la unidad.

4. Utilizar el gráfico de sedimentación y valorar donde está la máxima caída.

Gráfico de sedimentación



¿Cuál es el número óptimo de factores?

Modelo del AF

- Para el cálculo de las saturaciones (o las cargas factoriales L_i), recordar que son los coeficientes asociados a los factores.

- Las cargas se pueden ver como las correlaciones entre las variables (cuando estos se consideran ortogonales): $L_{ij} = Cor\ (X_i, Y_j)$

- Para conocer la porción de la variancia de la variable X_i explicada por el factor F_k, utilizamos la comunalidad, la cual se expresa como:

$Comm(X_j) = \sum_{i=1}^{m} \lambda_{ij}^2$ la cual se suele denotar como $Comm(X_j) = h_i^2$

Modelo del AF

- La variancia total para las variables se suele denotar como:

$$\mathrm{Var}(X_i) = 1 = \underbrace{\sum_{j=1}^{m} \lambda_{ij}^2}_{h_i^2 \,=\, \text{comunalidad}} + \underbrace{\psi_i}_{\mathrm{Var}(e_i) \,=\, \text{especificidad}}$$

- La descomponían total se escribe como:

$$\underbrace{\sum_{i=1}^{p} \mathrm{Var}(X_i)}_{\text{Variancia total}} = p = \underbrace{\sum_{i} \lambda_{i1}^2}_{\text{Variancia explicada por } Y_1 \;(=\lambda_1)} + \ldots + \underbrace{\sum_{i} \lambda_{im}^2}_{\text{Variancia explicada por par } Y_m \;(=\lambda_m)} + \underbrace{\sum_{i} \psi_i}_{\text{Variancia residual}}$$

La rotación factorial

- Las cargas factoriales obtenidas inicialmente se pueden transformar mediante una rotación ortogonal, para obtener nuevas cargas que todavía reproducen la matriz de correlación en forma equivalente.

- Existen también rotaciones oblicuas, es decir rotaciones de los ejes que no los mantienen perpendiculares.

- Para interpretar los factores corrientemente se agrupan las variables que tienen cargas altas en los mismos factores.

- Pero a veces las cargas factoriales originales no se pueden interpretar fácilmente debido a que la mayoría de los factores están correlacionados con muchas variables, y no ofrecen un patrón claro.

La rotación factorial

- Sin embargo, una rotación de las mismas puede facilitar esa interpretación mediante la búsqueda de una estructura simple de cargas.

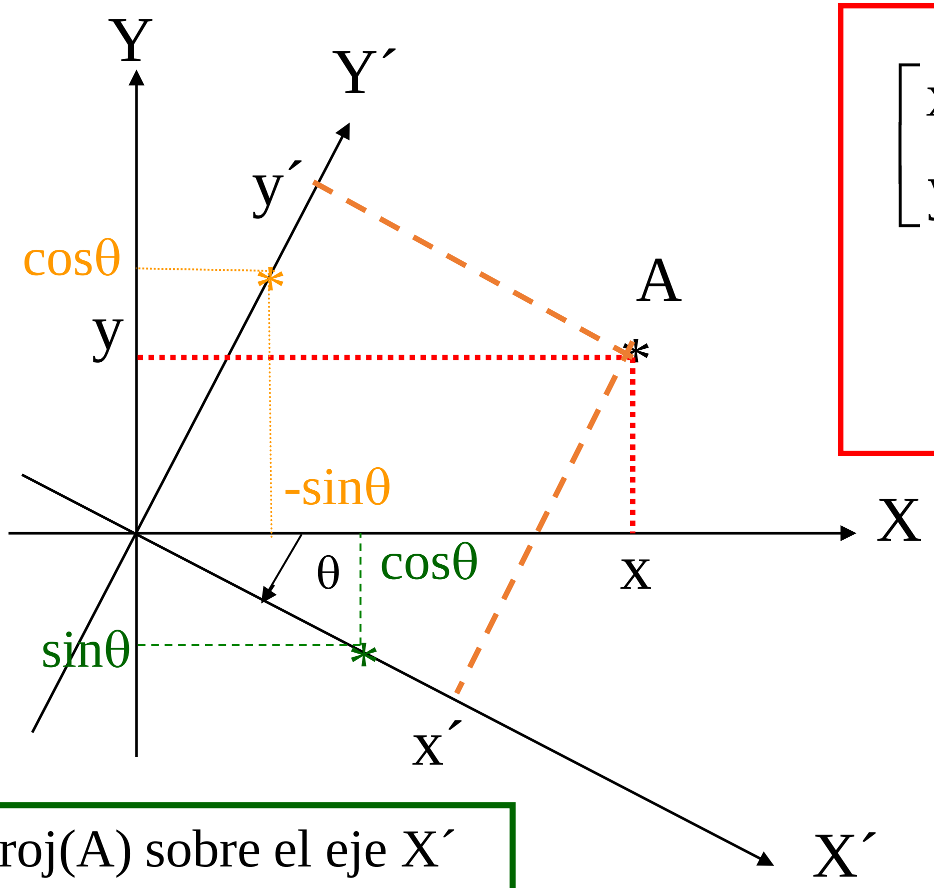
- Esta estructura más simple debería ser tal que cada factor tenga cargas no nulas en solo unas pocas variables, y que cada variable tenga cargas no nulas en solo unos cuantos factores.

- De esta manera cada factor se puede interpretar más fácilmente al estar asociado solamente a unas pocas variables.

- La rotación no afecta la bondad de ajuste de la solución factorial: la comunalidad y el % de variancia explicado por todos los factores no varían. Sin embargo, el porcentaje de variancia explicado individualmente por cada uno de los factores si cambia, produciendo una redistribución de la variancia total explicada entre los diferentes factores.

La idea de la rotación

Matriz de rotación de un ángulo θ



$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \underbrace{\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}}_{T} \begin{bmatrix} x \\ y \end{bmatrix}$$

Matriz de rotación T :

T´T = T T´= I

x´ = Proj(A) sobre el eje X´

y´ = Proj(A) sobre el eje Y´

La rotación factorial

- Hay diversos métodos de rotación, es decir, métodos para encontrar la matriz que rote la solución inicial y la transforme en una más fácil de intrepretar.

- El método más usado es el método VARIMAX. Intenta minimizar el número de variables que tienen cargas altas en un factor para así simplificar la interpretación de los factores.

- El Método QUARTIMAX busca minimizar el número de factores necesarios para explicar una variable, con la intención de simplificar la interpretación de las variables.

- El método EQUAMAX es una combinación del método VARIMAX y QUARTIMAX.

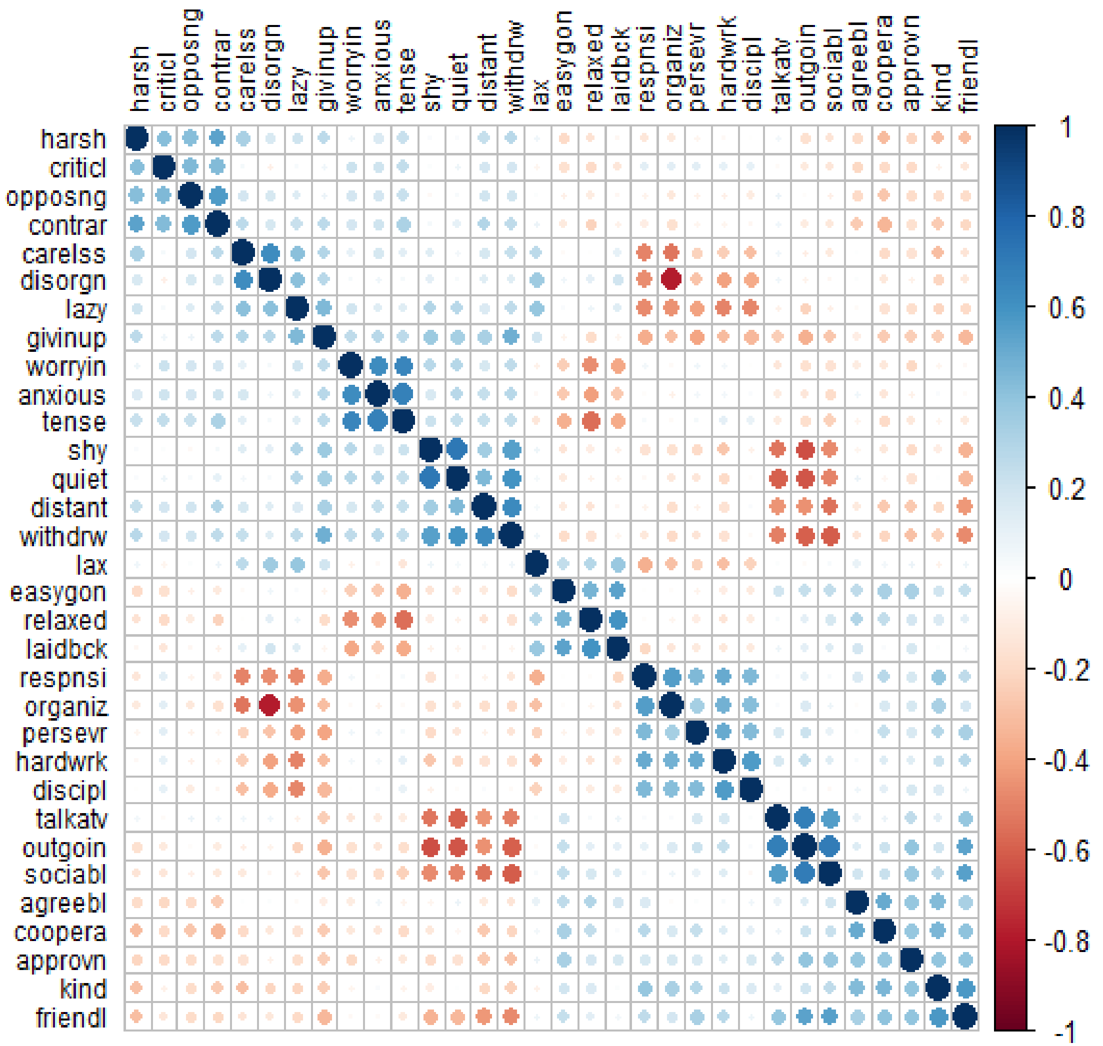
Etapas de un AF

- Al hacer un AF, se deberían llevar a cabo las siguientes etapas:

1. Análisis descriptivos de las variables.
2. Primera aproximación de la composición de los datos.
3. Determinar el número de las factores.
4. Segunda aproximación de la composición de los datos.
5. Ajustar diversos tipos de rotación.
6. Determinada la mejor solución en el agrupamiento de variables.

Análisis descriptivo de las variables

- Análisis de posición y variabilidad de cada variable.
- Distribución de las variables.
- Correlaciones y correlograma.

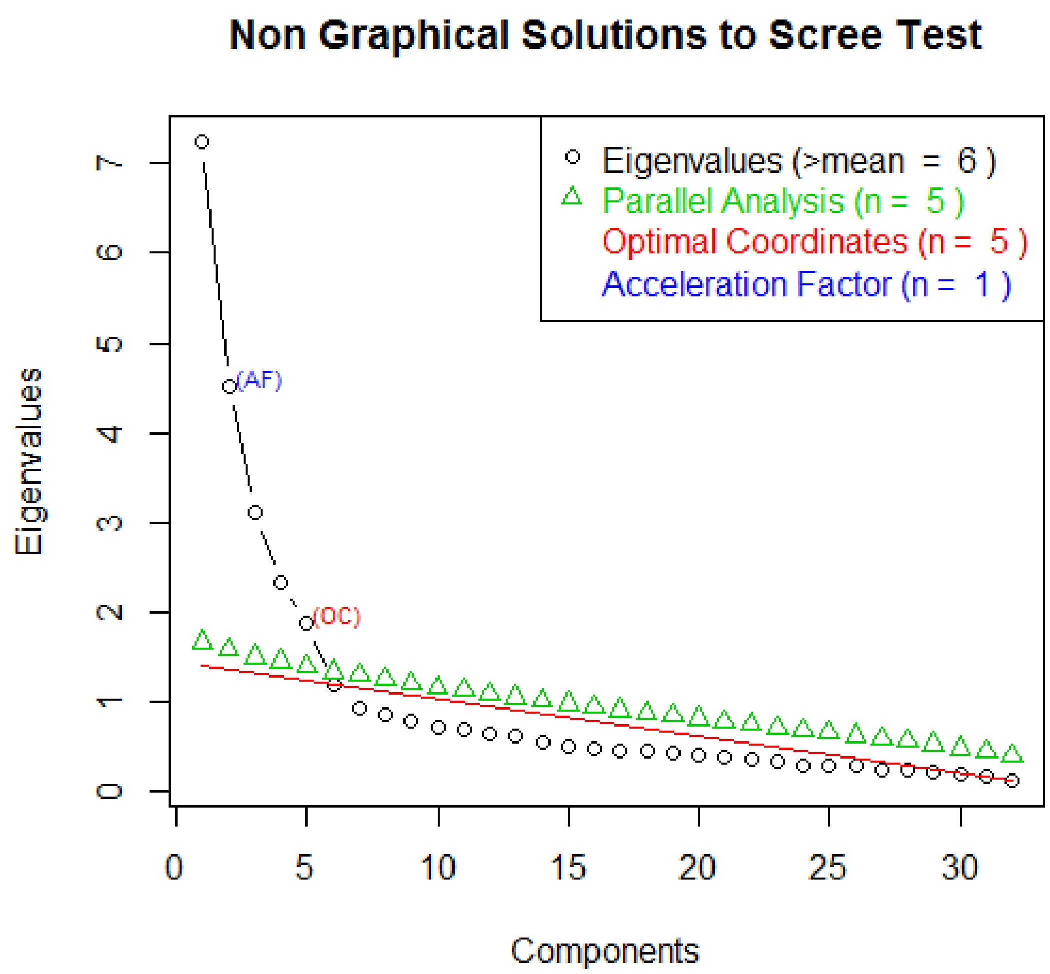


Primera aproximación de la estructura

- Se obtiene el primer modelo del AF.
- Se aconseja al inicio adecuar el modelo con un número relativamente grande de factores.
- Al inicio no introducimos rotaciones ni nada. Es el modelo simple, el ortogonal.
- Valoramos un poco qué variables son pertenecientes a qué factores.
- Esto nos permite ver si hay factores sobrantes.

Determinación del número de componentes

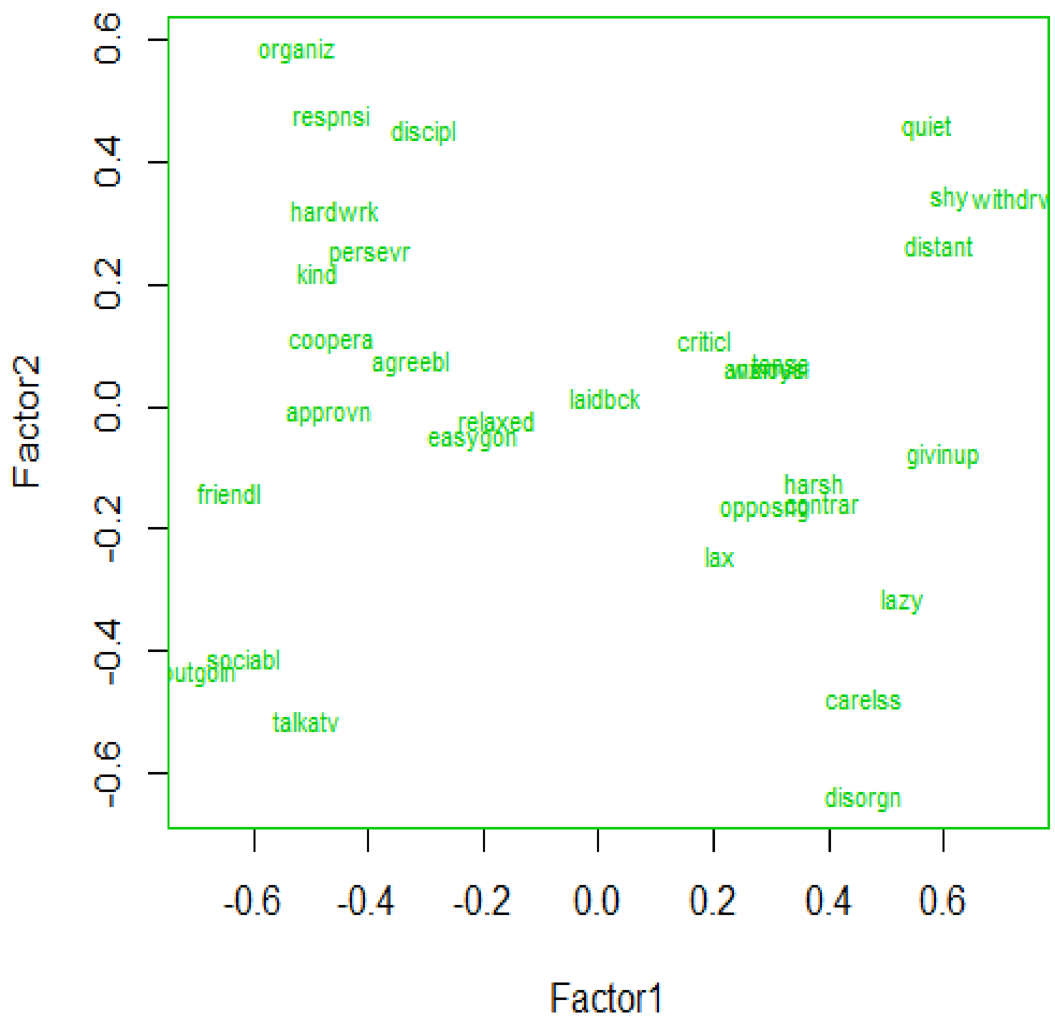
- Se debe determinar un número de factores.
- Hay diversos criterios, pero la mayoría de veces se prefiere utilizar el gráfico de sedimentación.
- El mismo gráfico de sedimentación posee otros criterios para saber cual podría ser la mejor estructura.



Agrupación de las variables

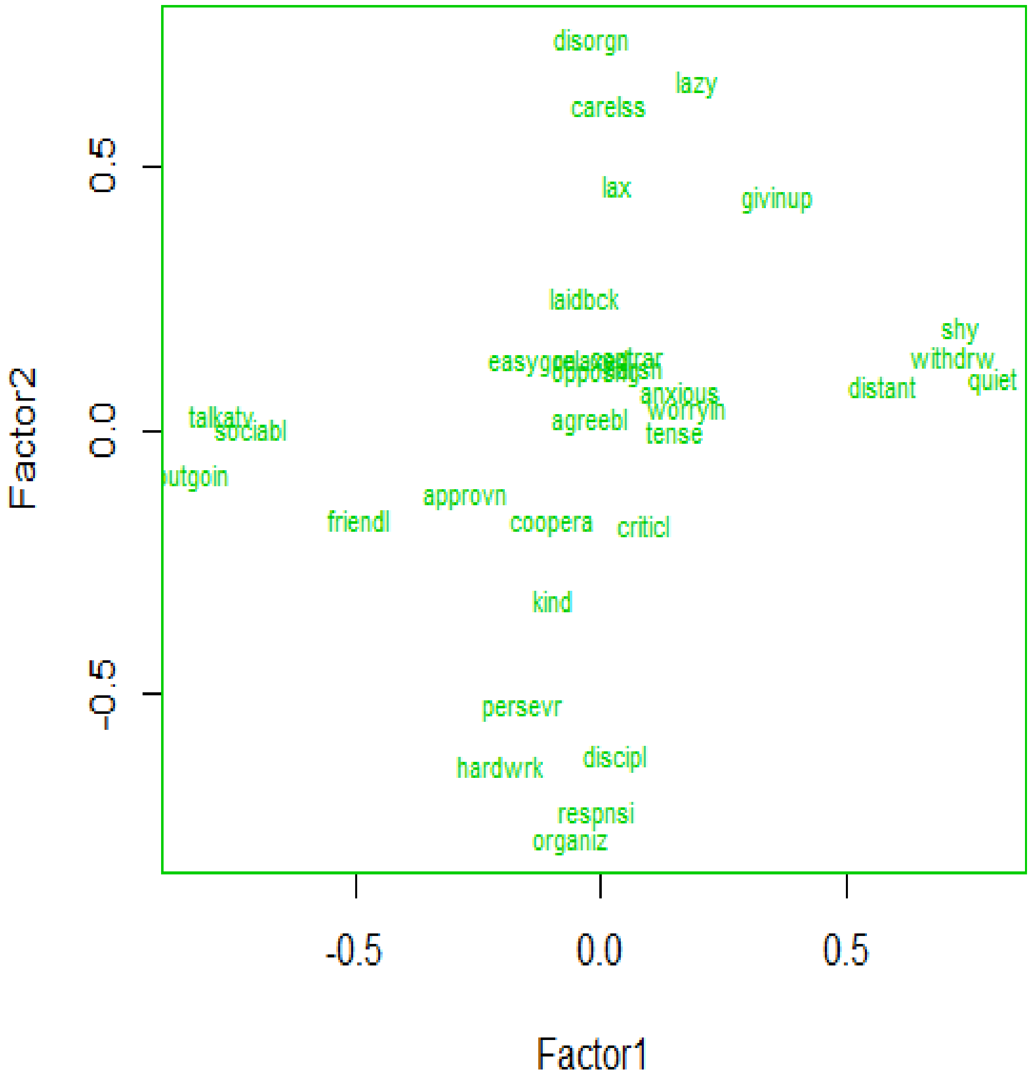
- Otra forma de ver la agrupación de las variables es mediante el gráfico de estas contra los factores.

- ¿Alguna idea de cuáles y cuántos grupos?



Fijación del n° de componentes + rotación

- Se vuelve a estimar el AF, pero esta vez ajustando por el número de factores y agregando un tipo de rotación.
- Esta parte suele ser bastante iterativa, dado que para ver valorar correctamente el nombre de los factores, se deban correr varios tipos de rotación.
- Una forma de ver el mejor ajuste, es corroborando de nuevo el gráfico de las variables contra los primeros factores.



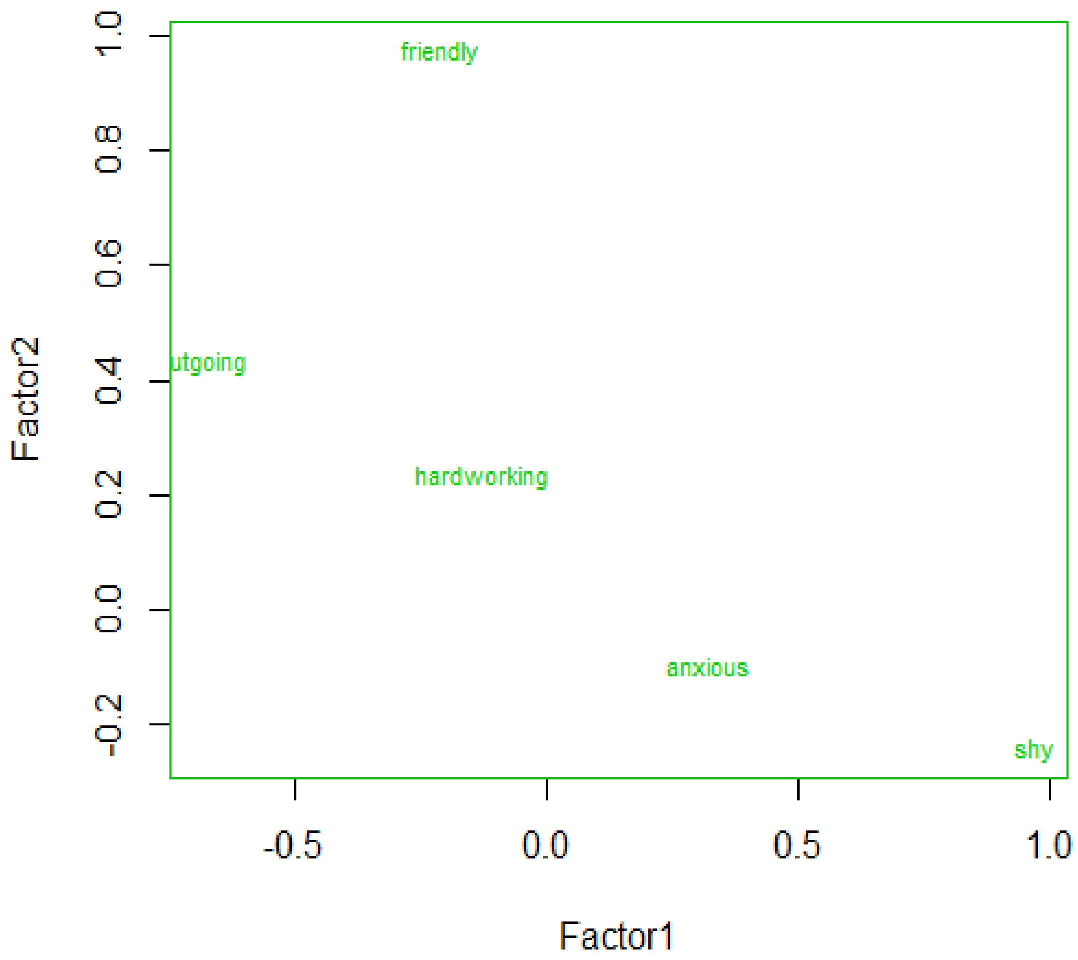
Agrupación final de las variables

- Agrupado el número de variables, conviene finalmente analizar si estas se juntaron correctamente.

- Es pertinente volver a correr el modelo del AF para el grupo de variables.

- Se busca que los grupos estén relativamente separados entre ellos.

- ¿Qué se puede finalmente concluir?



Y….¿qué pasó con los individuos?

- Aunque el AF se fundamente en la estructura subyacente de las variables, se pueden caracterizar a los individuos.

- En efecto, se suele utilizar el puntaje del análisis factorial (factor score), para saber cómo estos resultaron según cada uno de los factores.

- Puntajes altos en cierto factor indican que estos se representan fuertemente en dichos factores, y puntajes bajos lo contrario.

- Esto no lo veremos por ahora, pero en todo análisis se pueden caracterizar a los sujetos.

- Normalmente, todo el análisis sobre variables e individuos, lo podríamos replicar en el AF.

The End