Guía de trabajo en "R":

ANÁLISIS FACTORIAL Y ANÁLISIS DE COMPONENTES PRINCIPALES

Ayudantía Estadística IV (2015), Sociología, Universidad de Chile
Guía y procedimiento en "R" realizados por Raúl Zamora y Javier Esnaola
Profesor a cargo del ramo: Giorgio Boccardo

Sobre las técnicas

El AFC y el ACP forman parte de un conjunto de técnicas multivariadas denominadas de "interdependencia", ya que analizan la relación mutua entre un conjunto de variables.

Su finalidad principal, por lo tanto, no es el análisis de relaciones causales, sino la agrupación de variables, en función de la variabilidad que cada variable comparte con otras variables (varianza o covarianza).

Se busca la síntesis de la información proporcionada por cierto grupo de variables observadas, en un número inferior de variables no observadas (factores comunes o componentes principales, depende de la variedad analítica que se realice), con la menor pérdida de información posible. Dichas variables no observadas (o latentes), darían cuenta de conceptos no observables directamente, que engloban lo que tienen en común las variables observadas.

Esta serie menor de variables latentes (factores o componentes) se caracteriza por aglutinar variables empíricas que están bastante correlacionadas entre sí y escasamente correlacionadas con aquellas variables empíricas que conforman otra estructura latente (o dimensión del concepto que se analice).

Esto implica que la "no correlación" entre grupos de variables es una propiedad importante. Significa que los indicadores miden "dimensiones" diferentes en los datos.

La técnica nos es útil, por tanto, para una serie de objetivos:

- ✓ Reducir la información de una matriz de correlaciones a partir de la construcción de funciones lineales.
- ✓ Descifrar patrones de dependencia a partir del análisis de correlaciones múltiples.
- ✓ Identificar dimensiones que representen esquemas conceptuales de análisis.
- ✓ Validar la construcción de instrumentos de medida, particularmente escalas.

Sobre la diferencia entre ACP y AFC

El análisis de componentes principales (ACP) se caracteriza por analizar la varianza total del conjunto de variables observadas. De ellas trata de determinar las dimensiones básicas (o "componentes") que las definen. En el análisis de factor común (AFC) el estudio de las interrelaciones entre las variables se restringe, en cambio, a la varianza común (o covarianza), es decir, a la búsqueda de un número reducido de "factores" que expresen lo que es "común" al conjunto de variables observadas.

En ACP se considera la varianza total de la serie de variables observadas. El propósito es maximizar la proporción total de la varianza explicada. En cambio, AFC está orientado al análisis de la

covarianza (varianza en común o comunalidad), no de la varianza total. En esta última modalidad analítica, la varianza se descompone en varianza común (o comunalidad) y varianza específica. La comunalidad de cada variable (h^2) expresa la porción de la varianza total de la variable "x" que es compartida con las p-1 variables observadas restantes. La varianza específica es, por el contrario, la porción de la varianza total de la variable que no es explicada por los "factores comunes" (que está compuesta por la varianza específica y el error). En palabras simples, en el AFC los factores explican las variables, y en el ACP las variables explican los factores.

Supuestos de las técnicas

Los supuestos de la técnica son la normalidad univariante y multivariante[1], colinealidad (0,3 mínimo, al menos cada variable debe tener un par de variables que cumpla, idealmente para todo el par de variables) y multicolinealidad.

El ACP, y algunos procedimientos de AF (como ejes principales), si bien no requieren normalidad en sus variables para poder entrar en el análisis, la asimetría afecta mucho la conformación de factores/componentes. Lo ideal es siempre entrar a probar modelos con variables normales, o al menos, simétricas.

También, no como supuestos estadísticos, por como principio o requerimiento analítico, se busca que la solución factorial sea sencilla, compuesta por el menor número posible de factores o componentes (principios de parsimonia y simplicidad). Junto con ello, se requiere que los factores extraídos sean estadísticamente significativos, y susceptibles de interpretación sustantiva.

Sobre las variables

Optimo: variables métricas. En caso de utilizar ordinales que se asuman como métricas (ej.: ítems de una escala Likert), se toleran aquellas variables con 5 categorías o más. No es necesario que las variables tengan el mismo rango, ya que tanto la matriz de correlación que se utiliza para los cálculos, como los procedimientos en sí de ACP y AFC estandarizan las variables cuando trabajan con ellas.

Sobre el tamaño muestral

Mínimo 50 casos, sugerido más de 200. Al menos 10 casos por cada variable; la cantidad de variables no debe exceder la mitad de los casos.

Sobre los casos perdidos

Según tamaño muestra, analizar pertinencia de eliminar o imputar casos con datos perdidos.

En caso de eliminar casos perdidos, existen dos formas:

1) Eliminar casos según lista (listwise). Esta forma como primera opción. Si no se está bajo en los n muestrales, lo recomendable es trabajar de esta forma.
2) Eliminar casos según pareja (pairwise). Esta es la última opción, la menos recomendable.

En caso de imputar:

[1] Su no cumplimiento limita el uso de extracción de factores por Máxima Verosimilitud y Mínimos Cuadrados. Implica también la imposibilidad de interpretar el test de Bartlett, pero permite el resto de los análisis. Se sugiere que como mínimo cada variable tenga un s de asimetría de +-2 –si no, debe considerarse tratar o descartar la variable.

1) Reemplazar por la media de regresión. Tener esta forma de tratamiento como segunda opción.
2) Reemplazar por la media de todos los casos.

Recordar también que debe haber un tratamiento de casos aberrantes (outliers o atípicos). Esto se puede realizar una vez realizados las estimaciones con listwise.

Fuentes:
Cea D'Ancona, M.A. (2002). Análisis multivariante: teoría y práctica en la investigación social.
Apuntes y presentación de clases "Estadística IV" (2013), dictada por Prof. Gabriela Azócar.

Preparación de base y variables

Para comenzar, localizamos el directorio (carpeta) de trabajo, cargamos la base de datos en formato .csv, y categorizamos todas las puntuaciones 88 y 99 (No Sabe/No Responde) de todas las variables de la base de datos como datos perdidos.

```
setwd("C:/Users/Raúl/Desktop/Ayudantia AF/UDP")
#Recordar localizar la propia carpeta de trabajo según
#donde estén los archivos en el equipo en que estemos trabajando.

#Creamos la base a partir de un archivo .csv
base <- read.csv(file="UDP.csv", sep=";", dec=",")
summary(base)
#Nota: En este ejemplos las variables son ordinales de 4 categorías.
#Lo aceptado es mínimo 5.

#Categorizar como NA los 88 y 99 (NS/NR).
base[base==99] <- NA
base[base==88] <- NA
```

Para efectos de esta guía, no se revisará el supuesto de **normalidad univariante y multivariante**. Para revisar el procedimiento de dichas pruebas, ver la Guía Comprobación de supuestos en R.

Hay que recordar, sin embargo, que los procedimientos factoriales son particularmente sensibles a la asimetría. Incluso para los métodos de extracción que son tolerantes a la no normalidad (ACP y AFC por ejes principales), es problemático el hecho de que se viole el supuesto de simetría en más 2 puntos en el **S de simetría**. 2,5 puntos sería el máximo tolerable.

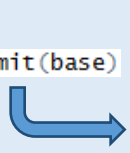
Supuesto de colinealidad

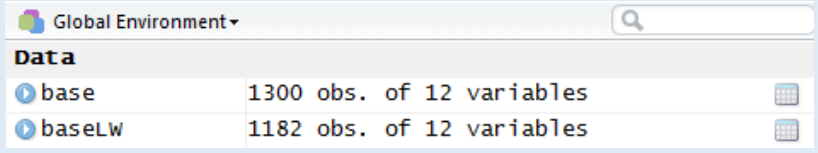
Para analizar la colinealidad existente entre las variables que se introducirán al modelo, y tener un panorama general sobre el estado de la multicolinealidad, podemos observar la **matriz de correlación**. En este caso, utilizaremos una matriz de correlación de Pearson. Generamos esta salida, junto con el **determinante de la matriz**, de la siguiente forma:

```
#Matriz de Correlaciones (denominado objeto "Rcor") y determinante de la matriz.
Rcor <- cor(base, use="pairwise.complete.obs")
Rcor
det(Rcor)
#Esta tabla permite evaluar la posibilidad de eliminar variables con baja
#colinealidad. Obviamente, la decisión no se toma exclusivamente aquí;
#lo correcto sería observar resultados de las pruebas subsecuentes y tomar
#una decisión a partir de toda la evidencia. Se espera que puedan justificar
#la inclusión/exclusión de variables, comparar modelos y revisar ajustes.
```

Nota: Recordar que esta matriz de correlación de Pearson, generada con los comandos de la imagen, está utilizando un método de tratamiento de NA por "pairwise". En caso de querer trabajar con "listwise", se elimina el use="pairwise.complete.obs", y se construye la matriz a partir de una base previamente ya generada de forma listwise, creada de la siguiente forma:

```
baseLW <- na.omit(base)
```





Lo mismo aplica para procedimientos y pruebas posteriores: si se trabaja por "pairwise" se debe especificar en la prueba misma, y si se trabaja con "listwise" se trabaja desde un comienzo con la base filtrada.

	ï..P21_1	P21_2	P21_3	P21_4	P21_5	P21_6	P21_7	P21_8	P21_9
ï..P21_1	1.0000000	0.5840791	0.4290672	0.2631741	0.2805185	0.2656951	0.2919217	0.3031242	0.2871116
P21_2	0.5840791	1.0000000	0.5442759	0.2381749	0.2744164	0.2304106	0.2896542	0.3324617	0.2600057
P21_3	0.4290672	0.5442759	1.0000000	0.2538186	0.2176172	0.1469773	0.2491463	0.2838129	0.2354281
P21_4	0.2631741	0.2381749	0.2538186	1.0000000	0.5905079	0.3727212	0.3416100	0.4115571	0.2230416
P21_5	0.2805185	0.2744164	0.2176172	0.5905079	1.0000000	0.6342718	0.4355488	0.4015818	0.1872870
P21_6	0.2656951	0.2304106	0.1469773	0.3727212	0.6342718	1.0000000	0.4939821	0.3379820	0.1830051
P21_7	0.2919217	0.2896542	0.2491463	0.3416100	0.4355488	0.4939821	1.0000000	0.4827069	0.2285462
P21_8	0.3031242	0.3324617	0.2838129	0.4115571	0.4015818	0.3379820	0.4827069	1.0000000	0.3206406
P21_9	0.2871116	0.2600057	0.2354281	0.2230416	0.1872870	0.1830051	0.2285462	0.3206406	1.0000000
P21_10	0.2418598	0.2727955	0.2731176	0.3905602	0.2891603	0.1949669	0.2444638	0.3666018	0.4170986
P21_11	0.1807132	0.2386152	0.3126559	0.3345583	0.2231140	0.1584647	0.2699344	0.3349098	0.3489761
P21_12	0.2501356	0.2607639	0.2759580	0.3122015	0.2723207	0.2311640	0.2558203	0.3407594	0.2760561

	P21_10	P21_11	P21_12
ï..P21_1	0.2418598	0.1807132	0.2501356
P21_2	0.2727955	0.2386152	0.2607639
P21_3	0.2731176	0.3126559	0.2759580
P21_4	0.3905602	0.3345583	0.3122015
P21_5	0.2891603	0.2231140	0.2723207
P21_6	0.1949669	0.1584647	0.2311640
P21_7	0.2444638	0.2699344	0.2558203
P21_8	0.3666018	0.3349098	0.3407594
P21_9	0.4170986	0.3489761	0.2760561
P21_10	1.0000000	0.6249645	0.3984890
P21_11	0.6249645	1.0000000	0.4752280
P21_12	0.3984890	0.4752280	1.0000000

En la matriz (creada como objeto *Rcor* en el programa) se pueden observar las correlaciones parciales, es decir, a nivel bivariado. Como buscamos colinealidad, esperamos que los valores fuera de la diagonal (los que no son la correlación de la variable con sí misma) sean mayores a 0,3 (baja colinealidad), más tendientes hacia 0,5 y más (colinealidad media), y óptimamente igual o mayor a 0,7 (colinealidad alta). Si alguna variable no cumple con colinealidad para/con todas las demás variables, se sugiere continuar trabajando, pero dejando anotado que para ese par no se cumple el supuesto. Bajas colinealidades pueden ser una razón de que los modelos en AFC y ACP no ajusten bien, por tanto la información que nos proporciona esta matriz se puede considerar para eliminar al final la variable si genera demasiados problemas para la buena convergencia.

En el caso de esta matriz, podemos observar que gran cantidad de ítems poseen bajas colinealidades con las demás ingresadas, y esto tiende a ser -en mayor y menor medida- la norma. Ítems particularmente problemáticos a este respecto podrían ser p21_11 y p21_6.

Junto con la matriz de correlaciones, generamos el determinante de esta:

```
> det(Rcor)
[1] 0.0161188
```

Un determinante bajo, es decir, cercano a 0, indica alta multicolinealidad entre las variables. No debe ser, sin embargo, igual a cero (matriz no singular), pues esto indicaría que algunas de las variables son linealmente dependientes y no se podrían realizar ciertos cálculos necesarios para los procedimientos multivariados. En este caso observamos que es muy cercano a 0, lo que sugiere alto nivel de colinealidad en el conjunto de variables involucradas en la matriz.

Un segundo procedimiento para el diagnóstico de colinealidad es la utilización de la **matriz de correlación anti-imagen**. Los comandos y salida en R son los siguientes:

```r
#Matriz anti-imagen (objeto A)
invRcor <- solve(Rcor)
A <- matrix(1,nrow(invRcor),ncol(invRcor))
for (i in 1:nrow(invRcor)){
  for (j in (i+1):ncol(invRcor)){
    A[i,j] <- invRcor[i,j]/sqrt(invRcor[i,i]*invRcor[j,j])
    A[j,i] <- A[i,j]
  }
}
colnames(A) <- colnames(base)
rownames(A) <- colnames(base)
print(A)
```

```
           ï..P21_1       P21_2       P21_3       P21_4       P21_5       P21_6       P21_7
ï..P21_1  1.00000000 -0.410377801 -0.136438002 -0.05070891 -0.006996510 -0.064681859 -0.050336539
P21_2    -0.41037780  1.000000000 -0.360753216  0.05071494 -0.050285428 -0.006642353 -0.031003528
P21_3    -0.13643800 -0.360753216  1.000000000 -0.05640549 -0.007118027  0.065209148 -0.042788810
P21_4    -0.05070891  0.050714937 -0.056405492  1.00000000 -0.421446033  0.036005836 -0.017401553
P21_5    -0.00699651 -0.050285428 -0.007118027 -0.42144603  1.000000000 -0.480576799 -0.059968746
P21_6    -0.06468186 -0.006642353  0.065209148  0.03600584 -0.480576799  1.000000000 -0.287037331
P21_7    -0.05033654 -0.031003528 -0.042788810 -0.01740155 -0.059968746 -0.287037331  1.000000000
P21_8    -0.01940942 -0.083462002 -0.019971784 -0.13043852 -0.071861754 -0.001374688 -0.284517018
P21_9    -0.12067502 -0.021338597 -0.017895412  0.01397962  0.037346541 -0.040228713 -0.018364557
P21_10   -0.02185006 -0.045958974  0.013700371 -0.13571290 -0.043174057  0.020111077  0.044505612
P21_11    0.08434884  0.009138033 -0.130394711 -0.06590064  0.045228062  0.033667138 -0.091356554
P21_12   -0.05473110 -0.021749043 -0.049960257 -0.04915635 -0.023850924 -0.050902116  0.003407644
              P21_8       P21_9      P21_10       P21_11       P21_12
ï..P21_1 -0.019409423 -0.12067502 -0.02185006  0.084348838 -0.054731102
P21_2    -0.083462002 -0.02133860 -0.04595897  0.009138033 -0.021749043
P21_3    -0.019971784 -0.01789541  0.01370037 -0.130394711 -0.049960257
P21_4    -0.130438523  0.01397962 -0.13571290 -0.065900640 -0.049156352
P21_5    -0.071861754  0.03734654 -0.04317406  0.045228062 -0.023850924
P21_6    -0.001374688 -0.04022871  0.02011108  0.033667138 -0.050902116
P21_7    -0.284517018 -0.01836456  0.04450561 -0.091356554  0.003407644
P21_8     1.000000000 -0.11443414 -0.07940700 -0.025539110 -0.096630547
P21_9    -0.114434139  1.00000000 -0.21406508 -0.076489617 -0.041390563
P21_10   -0.079407001 -0.21406508  1.00000000 -0.461984391 -0.060883411
P21_11   -0.025539110 -0.07648962 -0.46198439  1.000000000 -0.273182917
P21_12   -0.096630547 -0.04139056 -0.06088341 -0.273182917  1.000000000
```

Esta matriz muestra el negativo del coeficiente de las correlaciones parciales (aquellas que se estiman entre par de variables sin considerar el efecto de las demás). Se interpreta de forma inversa a la matriz de correlación: cuando los valores fuera de la diagonal son bajos (cercanos a 0), se está en presencia de alta colinealidad entre pares de variables.

Supuesto de multicolinealidad

Para el diagnóstico de la multicolinealidad de las variables que ingresaremos a los modelos, además de interpretar el **determinante de la matriz de correlaciones**, trabajaremos con el **test de esfericidad de Bartlett** y la prueba de **Kaiser-Meyer-Olkin** (**KMO** para los amigos).

El **test de esfericidad de Bartlett** busca contrastar la hipótesis nula de que la matriz de correlaciones es igual a una matriz de identidad[2]. Lo que nos interesa para efectos de buscar multicolinealidad, por lo tanto, es rechazar la hipótesis nula, y aceptar la hipótesis alternativa de que la matriz es distinta a una matriz de identidad, y por ende hay un nivel suficiente de multicolinealidad entre las variables. Este procedimiento es particularmente útil cuando el tamaño muestral es pequeño.

[2] En una matriz de identidad la diagonal es 1, y los valores fuera de la diagonal son 0. Esto implica que no hay más colinealidad entre las variables que la que hay entre cada variable consigo misma.

Es importante tener en cuenta que el test asume que existe normalidad multivariante en el conjunto de variables involucradas. Esto hace que en ausencia de dicho supuesto, la prueba no sea interpretable.

El procedimiento, para ser realizado en R, requiere instalar y cargar un paquete ("psych"). Los comandos en R para aquello, en conjunto con los comandos del test en sí, son los siguientes:

```
#Test de esfericidad de Bartlett.
install.packages("psych")
library(psych)
print(cortest.bartlett(Rcor,n = nrow(base)))
#El Chi2 da distinto en SPSS (5102,610). Esto no no es tan problemático.
#Los datos, si bien son distintos, son coherentes.
#Recordar que esta prueba sólo es interpretable si tenemos normalidad multivariante.
```

La salida de R es la siguiente:

```
> print(cortest.bartlett(Rcor,n = nrow(base)))
$chisq
[1] 5342.021

$p.value
[1] 0

$df
[1] 66
```

El test entrega un Chi2 y los grados de libertad, lo que permite generar un p valor. En este caso, al no contar con normalidad multivariante, no podemos interpretar la prueba. En caso de haber contado con el supuesto, esta misma salida se habría interpretado como presencia de multicolinealidad en el conjunto de variables, ya que el p valor es menor a 0,05.

Como última prueba de multicolinealidad antes de comenzar probar modelos, analizaremos el **KMO**. El índice KMO compara la magnitud de los coeficientes de correlación observados con la magnitud de los coeficientes de correlación parcial. Este estadístico varía entre 0 y 1, y se pueden calificar de la siguiente forma:

0,90 > KMO	Muy bueno
0,90 > KMO > 0,80	Bueno
0,80 > KMO > 0,70	Aceptable
0,70 > KMO > 0,60	Mediocre o regular
0,60 > KMO > 0,50	Malo
0,50 > KMO	Inaceptable o muy malo[3]

Los comandos y salidas de R para este ejemplo del procedimiento son los siguientes:

```
#KMO
kmo.num <- sum(Rcor^2) - sum(diag(Rcor^2))
kmo.denom <- kmo.num + (sum(A^2) - sum(diag(A^2)))
kmo <- kmo.num/kmo.denom
print(kmo)
```

```
> print(kmo)
[1] 0.8402373
```

Para este caso, el índice KMO nos indica un buen nivel de multicolinealidad entre las variables.

[3] Si KMO es menor a 0.5, hay que entrar a considerar cambiar de variables o de técnica, ya que es muy poco probable que funciones los modelos sin el cumplimiento de esta prueba.

Análisis de Componentes Principales en "R"

Con los supuestos revisados, podemos comenzar a trabajar con los procedimientos multivariados. Comenzaremos con ACP. Para ello, es necesario instalar y cargar un paquete ("GPArotation").

```
install.packages("GPArotation")
library(GPArotation)
```

Para comenzar, nos interesa explorar la cantidad de componentes a extraer. Para ello, solicitamos tantos componentes como variables involucradas en el análisis (12).

```
#ACP sin rotacion n°1 (explorando cantidad de componentes)
CPnorotado1 <- principal(base, nfactors=12, rotate="none", use=pairwise)
CPnorotado1
#Autovalor (SS loadings): medida de la varianza que explica cada componente/factor.
```

La salida que nos entrega R es la siguiente:

```
Principal Components Analysis
Call: principal(r = base, nfactors = 12, rotate = "none", use = pairwise)
Standardized loadings (pattern matrix) based upon correlation matrix
           PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11  PC12 h2       u2
ï..P21_1 0.59  0.14  0.56  0.05  0.15  0.17 -0.14 -0.36  0.20  0.29  0.04 -0.05  1  1.1e-16
P21_2    0.61  0.23  0.57 -0.06  0.01  0.00  0.03 -0.12 -0.19 -0.43  0.07  0.06  1  1.1e-15
P21_3    0.56  0.33  0.43 -0.24 -0.09 -0.18  0.20  0.45 -0.01  0.20 -0.10  0.00  1 -4.4e-16
P21_4    0.66 -0.26 -0.19 -0.33  0.31 -0.26 -0.23  0.05  0.28 -0.09  0.04  0.21  1  1.1e-15
P21_5    0.67 -0.54 -0.05 -0.18  0.22  0.05  0.02  0.07 -0.12 -0.03  0.02 -0.39  1  1.9e-15
P21_6    0.58 -0.60  0.02  0.08  0.03  0.32  0.25  0.03 -0.21  0.12  0.03  0.28  1  7.8e-16
P21_7    0.63 -0.35  0.04  0.29 -0.42 -0.11  0.21 -0.04  0.36 -0.13 -0.10 -0.06  1  1.4e-15
P21_8    0.68 -0.10 -0.06  0.24 -0.28 -0.38 -0.37 -0.04 -0.30  0.13  0.05  0.02  1  1.0e-15
P21_9    0.52  0.31 -0.14  0.61  0.36  0.12 -0.07  0.28  0.06 -0.06  0.04 -0.02  1 -1.1e-15
P21_10   0.64  0.34 -0.40 -0.03  0.17 -0.10  0.20 -0.27 -0.12  0.02 -0.37  0.01  1  6.7e-16
P21_11   0.61  0.40 -0.44 -0.13 -0.11 -0.02  0.27 -0.10  0.03  0.05  0.40 -0.03  1  1.1e-15
P21_12   0.59  0.23 -0.26 -0.22 -0.33  0.49 -0.33  0.12  0.04 -0.05 -0.10  0.00  1  3.3e-16

                       PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10 PC11 PC12
SS loadings           4.51 1.46 1.31 0.79 0.72 0.65 0.59 0.53 0.45 0.37 0.34 0.29
Proportion Var        0.38 0.12 0.11 0.07 0.06 0.05 0.05 0.04 0.04 0.03 0.03 0.02
Cumulative Var        0.38 0.50 0.61 0.67 0.73 0.79 0.84 0.88 0.92 0.95 0.98 1.00
Proportion Explained  0.38 0.12 0.11 0.07 0.06 0.05 0.05 0.04 0.04 0.03 0.03 0.02
Cumulative Proportion 0.38 0.50 0.61 0.67 0.73 0.79 0.84 0.88 0.92 0.95 0.98 1.00

Test of the hypothesis that 12 components are sufficient.

The degrees of freedom for the null model are  66  and the objective function was  4.13
The degrees of freedom for the model are -12  and the objective function was  0
The total number of observations was  1300  with MLE Chi Square =  0  with prob <  NA

Fit based upon off diagonal values = 1
```

Lo que interesa observar en este primer momento, son los **autovalores** *(SS loadings)* de los componentes extraídos (cuadrado rojo). Los autovalores nos darán una medida de tolerancia para poder decidir con cuanta cantidad de componentes es recomendable quedarnos. Autovalores iguales o mayores a 1 indican que el componente logra explicar más varianza que una variable por sí sola. A partir de los autovalores en este caso, podemos inclinarnos por trabajar con 3 componentes. Una solución con tres componentes, explicaría un 61% de la varianza total del conjunto de variables (círculo verde), lo que está casi justo en el límite de lo tolerable en ciencias sociales.

Por ahora, no interpretaremos la **matriz de componentes no rotados** (cuadrado azul), ya que nos interesa observar los puntajes lambda de los ítems sólo en los componentes con los que nos vamos a quedar.

Probaremos la solución con 3 componentes. Los comandos y salida en R para aquello son los siguientes:

```
#ACP sin rotacion n°2 (con cantidad de componentes decididos, para sacar matriz no rotada)
CPnorotado2 <- principal(base, nfactors=3, rotate="none", use=pairwise)
CPnorotado2
```

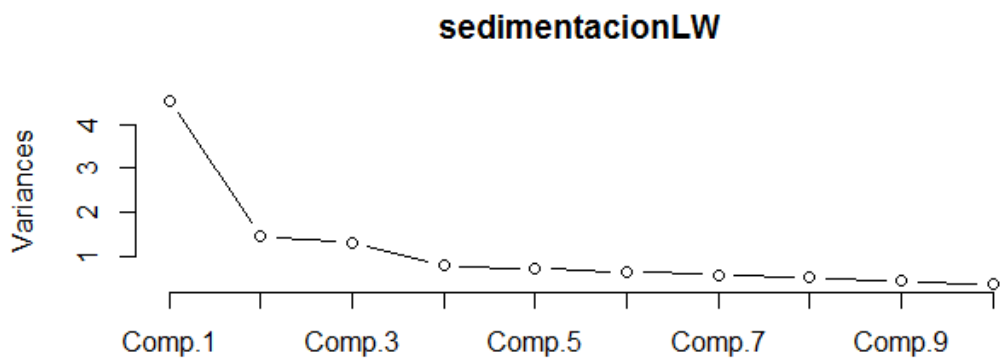
```
Principal Components Analysis
Call: principal(r = base, nfactors = 3, rotate = "none", use = pairwise)
Standardized loadings (pattern matrix) based upon correlation matrix
           PC1   PC2   PC3   h2   u2
i..P21_1 0.59  0.14  0.56 0.67 0.33
P21_2    0.61  0.23  0.57 0.75 0.25
P21_3    0.56  0.33  0.43 0.62 0.38
P21_4    0.66 -0.26 -0.19 0.54 0.46
P21_5    0.67 -0.54 -0.05 0.74 0.26
P21_6    0.58 -0.60  0.02 0.69 0.31
P21_7    0.63 -0.35  0.04 0.52 0.48
P21_8    0.68 -0.10 -0.06 0.48 0.52
P21_9    0.52  0.31 -0.14 0.39 0.61
P21_10   0.64  0.34 -0.40 0.70 0.30
P21_11   0.61  0.40 -0.44 0.72 0.28
P21_12   0.59  0.23 -0.26 0.47 0.53

                       PC1  PC2  PC3
SS loadings           4.51 1.46 1.31
Proportion Var        0.38 0.12 0.11
Cumulative Var        0.38 0.50 0.61
Proportion Explained  0.62 0.20 0.18
Cumulative Proportion 0.62 0.82 1.00

Test of the hypothesis that 3 components are sufficient.

The degrees of freedom for the null model are  66  and the objective function was  4.13
The degrees of freedom for the model are 33  and the objective function was  0.68
The total number of observations was  1300  with MLE Chi Square =  872.36  with prob <  6.1e-162

Fit based upon off diagonal values = 0.95
```

A partir de la matriz de componentes no rotados, observamos los **lambdas** de cada ítem sobre cada componente (cuadrado amarillo). El lambda nos da una medida de cuánto pesa cada ítem (variable observada) en relación al total de varianza que extractan. Forma parte de cada uno de los componentes generados (variables latentes). Estos valores van de 0 a 1.

Como ya revisamos en la solución anterior, los autovalores de los tres componentes son todos mayores a 1, y el modelo explica un 61% de la varianza total. Es el primer componente el que se lleva gran parte de la varianza.

Otro elemento que aporta a la delimitación de cantidad de factores/componentes a elegir, es el gráfico de sedimentación. Los comandos y la salida de R son los siguientes:

```
#Gráfico de sedimentación
sedimentacion <- princomp(base, scores=TRUE, cor=TRUE)
plot(sedimentacion, type="lines")
#El comando no arroja gráfico de sedimentación cuando se trabaja con una base incompleta
#(nuestra base posee NA). Para esta prueba en R, nos vemos forzados a trabajar
#con la base listwise ("baseLW"), que no posee NA ya que todos los casos tienen puntuación
#en todas las variables. SPSS si procesa los NA.
sedimentacionLW <- princomp(baseLW, scores=TRUE, cor=TRUE)
plot(sedimentacionLW, type="lines")
```



El gráfico de sedimentación muestra en el eje "y" los autovalores, y en el eje "x" posiciona los componentes. La disposición gráfica –particularmente los cambios en la pendiente– ayudan a observar cuanta capacidad explicativa va aportando cada componente a medida que se van incorporando al modelo.

Ahora probaremos la solución de tres componentes, pero con una **solución rotada** mediante el procedimiento de varianza máxima (Varimax para los amigos). Varimax es un ajuste de rotación de los componentes (en este caso, al ser ACP), que maximiza la varianza explicada por cada uno de ellos, equilibrando así las diferencias entre autovalores. Es por este motivo que muchas veces se considera que este tipo de rotación exagera la cantidad de componentes/factores que son apropiados de considerar, toda vez que se busca escoger un modelo parsimonioso[4]. La sintaxis y salidas de R para la solución rotada son las siguientes:

```
#ACP con rotación Varimax
CProtado <- principal(base, nfactors=3, rotate="varimax")
CProtado
#Cuando se rota por varimax, opera automaticamente pairwise.
#Esto no es problema, ya que si se requiere listwise, se trabaja
#con la base desde el principio así (en este caso, con objeto "baseLW").
#Existen distintos métodos de rotación. No profundizaremos mayormente en esto.
```

```
Principal Components Analysis
Call: principal(r = base, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
           RC2  RC1  RC3   h2   u2
ï..P21_1 0.23 0.10 0.78 0.67 0.33
P21_2    0.17 0.15 0.84 0.75 0.25
P21_3    0.07 0.26 0.74 0.62 0.38
P21_4    0.63 0.36 0.08 0.54 0.46
P21_5    0.84 0.13 0.11 0.74 0.26
P21_6    0.83 0.01 0.10 0.69 0.31
P21_7    0.67 0.16 0.22 0.52 0.48
P21_8    0.52 0.39 0.25 0.48 0.52
P21_9    0.10 0.56 0.25 0.39 0.61
P21_10   0.17 0.81 0.11 0.70 0.30
P21_11   0.11 0.84 0.08 0.72 0.28
P21_12   0.21 0.63 0.16 0.47 0.53
```

[4] Existen otros procedimientos de rotación, como por ejemplo "Oblimin", que es un tipo de rotación oblicua, que se utiliza cuando se considera que los componentes/factores a extraer no son completamente independientes entre sí, debido a que se entienden como pertenecientes a un mismo concepto latente general.

```
                       RC2  RC1  RC3
SS loadings           2.68 2.50 2.10
Proportion Var        0.22 0.21 0.17
Cumulative Var        0.22 0.43 0.61
Proportion Explained  0.37 0.34 0.29
Cumulative Proportion 0.37 0.71 1.00

Test of the hypothesis that 3 components are sufficient.

The degrees of freedom for the null model are  66  and the objective function was  4.13
The degrees of freedom for the model are 33  and the objective function was  0.68
The total number of observations was  1300  with MLE Chi Square =  872.36  with prob <  6.1e-16
2

Fit based upon off diagonal values = 0.95
```

Con la rotación Varimax, podemos observar cambios tanto en la matriz de componentes (ahora rotada), como en los autovalores.

Se observa a partir de la matriz de componentes rotados que los lambdas logran diferenciarse mejor, pudiendo identificarse fácilmente a qué variable latente (componente) tiende a asociarse cada ítem (variables observadas). Lambdas bien definidos nos pueden dar luces sobre un posible modelo confirmatorio.

También se observa que el modelo sigue explicando un 61% de la varianza total (como ya mencionamos, casi justo en el límite de lo tolerable en ciencias sociales), pero esta se maximizó para cada uno de los componentes, equilibrándose los autovalores. Con este cambio en los autovalores, podemos pensar en una solución rotada por Varimax que integre mayor cantidad de componentes. Para efectos de esta guía, no seguiremos buscando un mejor modelo.

Análisis de Factor Común en "R"

Para el análisis de factor común, trabajaremos con el **método de extracción** por ejes principales ("principal axis" en inglés, y "pa" para R), ya que no contamos con normalidad multivariante.

De forma exploratoria, generaremos una solución factorial no rotada, por ejes principales, y con 5 factores[5]. Los comandos y salidas en R son las que siguen.

```
#AFC
#En este caso, trabajaremos con método de extracción por ejes principales
#(principal axis, pa), ya que no contamos con normalidad multivariante.
#Para ello agregamos el "fm=pa" en la sintaxis.
AFnorotado1 <- fa(base, nfactors=5, fm="pa", rotate="none", max.iter = 100)
AFnorotado1
#promedio de h2 da 0,58.
#El modelo explica 58% de la varianza.

Factor Analysis using method =  pa
Call: fa(r = base, nfactors = 5, rotate = "none", max.iter = 100, fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
          PA1   PA2   PA3   PA4   PA5   h2     u2 com
ï..P21_1 0.55  0.11  0.43 -0.05  0.10 0.51  0.493 2.1
P21_2    0.60  0.21  0.55 -0.08 -0.07 0.71  0.290 2.3
P21_3    0.52  0.24  0.31 -0.06 -0.14 0.45  0.554 2.4
P21_4    0.60 -0.16 -0.15 -0.15  0.00 0.43  0.569 1.4
P21_5    0.71 -0.60 -0.11 -0.36 -0.05 1.01 -0.013 2.5
P21_6    0.54 -0.41 -0.01  0.02  0.04 0.47  0.533 1.9
P21_7    0.64 -0.31  0.03  0.58 -0.07 0.86  0.142 2.5
P21_8    0.62 -0.04 -0.03  0.13  0.09 0.41  0.589 1.1
P21_9    0.47  0.22 -0.07  0.04  0.32 0.38  0.622 2.3
P21_10   0.61  0.32 -0.31 -0.06  0.10 0.59  0.410 2.2
P21_11   0.61  0.43 -0.42  0.02 -0.21 0.77  0.228 3.0
P21_12   0.52  0.18 -0.15 -0.01 -0.04 0.33  0.672 1.4

                       PA1  PA2  PA3  PA4  PA5
SS loadings           4.11 1.15 0.91 0.53 0.21
Proportion Var        0.34 0.10 0.08 0.04 0.02
Cumulative Var        0.34 0.44 0.51 0.56 0.58
Proportion Explained  0.60 0.17 0.13 0.08 0.03
Cumulative Proportion 0.60 0.76 0.89 0.97 1.00
```

Lo primero que llama la atención, es la baja varianza que logra explicar el modelo, incluso con una alta cantidad de factores. Tenemos que con 5 factores (de 12 variables originalmente) logramos explicar sólo un 58% de la varianza. El modelo a este nivel ya es muy pobre en su capacidad explicativa, y lo será más aún a medida que tomamos la decisión de restringir la cantidad de factores a considerar.

A partir de los autovalores, podemos pensar a priori que sería razonable trabajar con una solución de 2 factores (ya que sólo 2 factores logran autovalor > 1). Existe, sin embargo, y a diferencia del ACP, otro indicador más preciso que ayuda a tomar la decisión a nivel estadístico. Este indicador es el **promedio de las comunalidades individuales** (promedio de la columna h^2, que indica para cada variable la comunalidad de cada una de estas con el conjunto de variables restantes). El promedio de las comunalidades individuales funciona como límite que indica con cuántos factores es razonable quedarse. En este caso, al ser 0.58, resulta razonable una solución de 3 factores.

[5] A diferencia del ACP, no pedimos modelos con tantos factores como variables para una primera exploración, ya que dicha solución en el caso de factorial no converge. Esto se debe a que a diferencia del ACP, AFC trabaja sólo con la covarianza, y no con la varianza total.

Cabe mencionar que para el caso de AFC, las comunalidades individuales (h^2) también son interpretables por sí solas. Muestran cómo covaría la variable individual con el resto de las variables. Ítems con comunalidad baja (<0,3) son problemáticos para la buena convergencia, parsimonia, e interpretabilidad de los modelos. A partir de las comunalidades podríamos llegar a justificar la exclusión de ciertos ítems problemáticos de nuestro modelo final[6].

Ahora observemos los comandos y salidas para una solución de 3 factores, aún sin rotar:

```
AFnorotado3 <- fa(base, nfactors=3, fm="pa", rotate="none", max.iter = 100)
AFnorotado3
#promedio de h2 da 0,51.
#El modelo explica 49% de la varianza.

 Factor Analysis using method =  pa
 Call: fa(r = base, nfactors = 3, rotate = "none", max.iter = 100, fm = "pa")
 Standardized loadings (pattern matrix) based upon correlation matrix
           PA1   PA2   PA3   h2   u2 com
 ï..P21_1 0.55  0.10  0.42 0.49 0.51 1.9
 P21_2    0.60  0.21  0.54 0.70 0.30 2.2
 P21_3    0.52  0.24  0.30 0.42 0.58 2.1
 P21_4    0.61 -0.19 -0.15 0.43 0.57 1.3
 P21_5    0.66 -0.51 -0.07 0.70 0.30 1.9
 P21_6    0.56 -0.49 -0.01 0.55 0.45 2.0
 P21_7    0.58 -0.24  0.01 0.39 0.61 1.3
 P21_8    0.62 -0.06 -0.04 0.39 0.61 1.0
 P21_9    0.46  0.19 -0.07 0.26 0.74 1.4
 P21_10   0.62  0.31 -0.34 0.59 0.41 2.1
 P21_11   0.60  0.38 -0.39 0.66 0.34 2.5
 P21_12   0.53  0.16 -0.16 0.33 0.67 1.4

                        PA1  PA2  PA3
 SS loadings           4.02 1.02 0.88
 Proportion Var        0.33 0.08 0.07
 Cumulative Var        0.33 0.42 0.49
 Proportion Explained  0.68 0.17 0.15
 Cumulative Proportion 0.68 0.85 1.00
```

Para esta solución, a la igual que la anterior, sigue siendo problemática la poca varianza total explicada (sólo un 49% en este caso). Respecto a los autovalores, los 3 logran ser mayores que el promedio de las comunalidades individuales (h^2), que es de 0.51, y sólo 2 alcanzan a ser mayores a 1.

Para el caso de AFC, también podemos buscar una solución comparando modelos rotados y no rotados. Al igual que en ACP, si trabajamos con rotación Varimax, al equilibrarse la varianza explicada por cada factor es posible pensar en ajustes con mayor cantidad de factores. Para efectos de esta guía, no seguiremos buscando más ajustes.

Para finalizar, una recomendación:

[6] Cuando la covarianza (comunalidad) es similar a la varianza total (cercana a 1) se espera que los modelos sean similares en sus resultados (AFC y ACP). En casos en que son muy distintas, por ejemplo comunalidades de 0,7 o menos, probablemente ambos modelos darán resultados diferentes (en este modelo la varianza específica es alta lo que significa que una proporción importante de la variación de la variable se explica por sí misma y no por la totalidad). Nótese que al cambiar el número de factores, varía la cantidad de covarianza total que el modelo utiliza, de ahí que las comunalidades parciales varíen cuando se cambia el número de factores que el modelo estima.

Es importante siempre tener en cuenta, tanto para el ACP como para el AFC, que la decisión de con cuántos componentes/factores quedarse no sólo debe tomarse a partir de criterios estadísticos, es decir, sólo en consideración de los autovalores (*autovalor > 1* en caso de ACP; *autovalor > 1* y *autovalor > promedio de* h^2 en caso de AFC), la varianza total explicada por el modelo, el método de extracción factores y la forma de rotación. También, la decisión debe tomarse a partir de criterios de interpretabilidad sociológica, coherencia teórica, simplicidad analítica, y parsimonia el modelo. Un buen trabajo es aquel que analiza varias soluciones posibles, ponderando y discutiendo todos estos aspectos (estadísticos y analíticos) al decidir con cual/cuales solución/es quedarse. En este momento también puede ser razonable eliminar casos atípicos, normalizar variables muy asimétricas, o eliminar variables problemáticas para tratar de buscar mejores ajustes.