

Métodos cuantitativos II: Análisis factorial exploratorio

Jose M Sallan
`jose.maria.sallan@upc.edu`

15 de febrero de 2018

Introducción al análisis factorial

Análisis factorial exploratorio

Un ejemplo de análisis factorial exploratorio

Introducción al análisis factorial

Análisis factorial exploratorio

Un ejemplo de análisis factorial exploratorio

El objetivo del análisis factorial es explicar la variabilidad de un conjunto de variables aleatorias observables en términos de un número menor de variables no observables (latentes) llamadas **factores**

Análisis factorial exploratorio (EFA)

- ▶ No hay ninguna hipótesis *a priori* sobre la cantidad y naturaleza de los factores
- ▶ Los factores se interpretan *a posteriori*, examinando qué variables observables correlacionan con cada factor

Análisis factorial confirmatorio (CFA)

- ▶ Se examina el ajuste de un modelo en el que hay hipótesis *a priori* acerca de qué variables están asociadas a cada factor
- ▶ CFA se lleva a cabo usando **modelos de ecuaciones estructurales**

Dos aproximaciones matemáticas:

Modelo de componentes

- ▶ variables observadas sin error
- ▶ estimación mediante componentes principales

Modelos de factor común

- ▶ variables observadas con error
- ▶ estimación mediante ejes principales, máxima verosimilitud

Matriz de correlación de calificaciones de estudiantes en seis materias:

	Gaélico	Inglés	Historia	Aritmética	Álgebra	Geometría
Gaélico	1.000	0.439	0.410	0.288	0.329	0.248
Inglés	0.439	1.000	0.351	0.354	0.320	0.329
Historia	0.410	0.351	1.000	0.164	0.190	0.181
Aritmética	0.288	0.354	0.164	1.000	0.595	0.470
Álgebra	0.329	0.320	0.190	0.595	1.000	0.464
Geometría	0.248	0.329	0.181	0.470	0.464	1.000

¿Cómo agruparían las materias? ¿Cómo son las correlaciones entre materias del mismo grupo?

Modelo de componentes con dos factores usando la función principal del paquete psych:

```
> library(psych)
> pr.scores <- principal(r=scores.cor, nfactors=2,
+                          rotate="none")
```


Modelo de factor común de dos factores usando la función `fa` del paquete `psych`, con `fm="pa"` para ejes principales y `fm="ml"` para máxima verosimilitud

```
> library(psych)
> pa.scores <- fa(r=scores.cor, nfactors=2, fm="pa",
+               rotate="none")
> ml.scores <- fa(r=scores.cor, nfactors=2, fm="ml",
+               rotate="none")
```

- ▶ Las **cargas factoriales** son la correlación entre variable observable y factor
- ▶ Si una variable i tiene una carga factorial alta en un factor j significa que gran parte de la variabilidad de i puede explicarse por el factor j
- ▶ Cada variable observable debería tener un valor de carga elevado en un solo factor

Cargas para el modelo de componentes principales (ejemplo de calificaciones de estudiantes):

```
> pr.scores$loadings
```

Loadings:

	PC1	PC2
Gaélico	0.658	0.449
Inglés	0.688	0.290
Historia	0.517	0.637
Aritmética	0.738	-0.413
Álgebra	0.744	-0.375
Geometría	0.678	-0.355

	PC1	PC2
SS loadings	2.733	1.130
Proportion Var	0.455	0.188
Cumulative Var	0.455	0.644

Los valores de las cargas factoriales no son únicos: cualquier **rotación** de los factores originales es también una solución válida

Una buena rotación relaciona cada variable con un solo factor

Algunos métodos de rotación: There are several rotation methods that try to attach each variable to a single factor:

- ▶ Ortogonales: (**varimax**, **quartimax**): mantienen los factores incorrelacionados (independientes)
- ▶ No ortogonal (**oblimin**)

La rotación por defecto en comandos de **psych** es oblimin

Ejemplo: calificaciones de estudiantes

Definiendo rotaciones

```
> pr.scores.varimax <- principal(r=scores.cor, nfactors=2,  
+                               rotate="varimax")  
> pa.scores.oblimin <- fa(r=scores.cor, nfactors=2,  
+                          fm="pa")  
> ml.scores.quartimax <- fa(r=scores.cor, nfactors=2, fm="ml",  
+                            rotate="quartimax")
```

Ejemplo: calificaciones de estudiantes

Solución rotada

```
> pr.scores.varimax$loadings
```

Loadings:

	RC1	RC2
Gaélico	0.223	0.765
Inglés	0.348	0.661
Historia		0.821
Aritmética	0.833	0.150
Álgebra	0.813	0.182
Geometría	0.749	0.156

	RC1	RC2
SS loadings	2.087	1.776
Proportion Var	0.348	0.296
Cumulative Var	0.348	0.644

```
> print(pr.scores.varimax$loadings, cutoff=0.4)
```

Loadings:

	RC1	RC2
Gaélico		0.765
Inglés		0.661
Historia		0.821
Aritmética	0.833	
Álgebra	0.813	
Geometría	0.749	

	RC1	RC2
SS loadings	2.087	1.776
Proportion Var	0.348	0.296
Cumulative Var	0.348	0.644

Introducción al análisis factorial

Análisis factorial exploratorio

Un ejemplo de análisis factorial exploratorio

150 observaciones de dos escalas de seis ítemes midiendo los focos regulatorios de promoción y prevención:

- ▶ **Foco de promoción:** completar tareas poniendo los medios necesarios para realizarlas con éxito (cumplir objetivos)
- ▶ **Foco de prevención:** completar tareas poniendo los medios necesarios para evitar aquello que impida su ejecución (evitar errores)

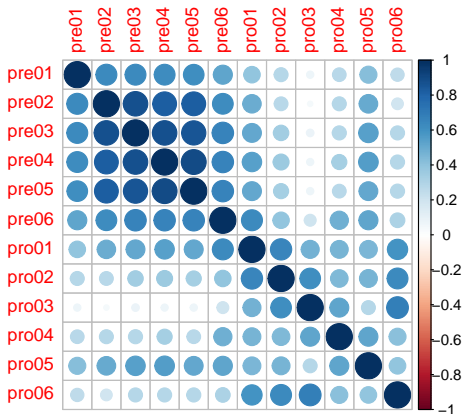
Obteniendo los datos y calculando la matriz de correlaciones:

```
> data <- read.csv("datascale.csv")  
> vars <- names(data)  
> focus <- data[  
+   ,which(grepl("pre", vars) | grepl("pro", vars))]  
> cor.focus <- cor(focus)
```

Ejemplo: escalas de foco regulatorio

Correlograma de variables

```
> library(corrplot)  
> corrplot(cor.focus, method="circle")
```



Antes de realizar un EFA, se ha de comprobar si las variables están correlacionadas:

- ▶ **Test de adecuación de la muestra de Kaiser-Meyer-Olkin:** evalúa si las correlaciones parciales son lo bastante grandes (debe ser mayor de 0,5)
- ▶ **Test de esfericidad de Bartlett:** examina la hipótesis nula de que la matriz de correlaciones es igual a la identidad

Ejemplo: escalas de foco regulatorio

Test de correlaciones no nulas

```
> library(psych)
> KMO(cor.focus)

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor.focus)
Overall MSA = 0.89
MSA for each item =
pre01 pre02 pre03 pre04 pre05 pre06 pro01 pro02 pro03 pro04 pro05 pro06
0.97 0.90 0.90 0.90 0.90 0.91 0.88 0.89 0.75 0.83 0.93 0.83

> cortest.bartlett(cor.focus, n=150)

$chisq
[1] 1429.374

$p.value
[1] 3.522197e-255

$df
[1] 66
```

¿Cuántos factores a extraer?

- ▶ Consideraciones teóricas: en este caso, la existencia de dos focos regulatorios
- ▶ Para el modelo de componentes: valores propios mayores que uno
- ▶ Proporción de varianza explicada (alrededor de 0,85)
- ▶ Análisis de sedimentación: función `fa.parallel`

Ejemplo: escalas de foco regulatorio

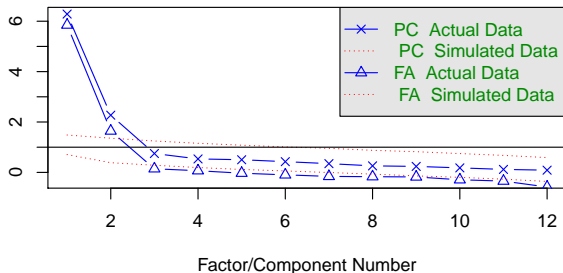
Número de factores a extraer

```
> fa.parallel(cor.focus, n.obs=150)
```

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

enevalues of principal components and factor an

Parallel Analysis Scree Plots



Ejemplo: escalas de foco regulatorio

Extracción del modelo

Las puntuaciones en un ítem de escala son variables con error de medida, así que es preferible usar un modelo de factor común (ejes principales con dos factores):

```
> pa.data.oblimin <- fa(r=cor.focus, nfactors=2, fm="pa")
```


Resultados del modelo

Algunos valores relevantes

- ▶ **h²** (communalities): variabilidad explicada por el análisis
- ▶ **u²** (uniquenesses): variabilidad explicada por el análisis
- ▶ **Proportion Var**: proporción de variabilidad explicada por el factor
- ▶ **Cumulative Var**: varianza explicada acumulada

Ejemplo: escalas de foco regulatorio

Resultados del modelo

```
> pa.data.oblimin  
Factor Analysis using method = pa  
Call: fa(r = cor.focus, nfactors = 2, fm = "pa")  
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PA1	PA2	h2	u2	com
pre01	0.67	0.04	0.47	0.53	1.0
pre02	0.93	-0.08	0.82	0.18	1.0
pre03	0.94	-0.02	0.87	0.13	1.0
pre04	0.93	0.00	0.87	0.13	1.0
pre05	0.92	-0.02	0.84	0.16	1.0
pre06	0.66	0.23	0.59	0.41	1.2
pro01	0.40	0.58	0.64	0.36	1.8
pro02	0.14	0.74	0.63	0.37	1.1
pro03	-0.21	0.90	0.73	0.27	1.1
pro04	0.18	0.55	0.41	0.59	1.2
pro05	0.48	0.34	0.46	0.54	1.8
pro06	0.04	0.77	0.62	0.38	1.0

	PA1	PA2
SS loadings	5.01	2.94
Proportion Var	0.42	0.24
Cumulative Var	0.42	0.66
Proportion Explained	0.63	0.37
Cumulative Proportion	0.63	1.00

```
With factor correlations of  
PA1 PA2  
PA1 1.00 0.33  
PA2 0.33 1.00
```

```
Mean item complexity = 1.2  
Test of the hypothesis that 2 factors are sufficient.
```

- ▶ El significado de los factores tiene que ver con el denominador común de las variables con carga factorial elevada en el factor
- ▶ Ejemplo de las capacidades de los estudiantes: capacidades verbales vs numéricas

Ejemplo: escalas de foco regulatorio

Interpretación de los factores

```
> print(pa.data.oblimin$loadings, cutoff=0.4)
```

Loadings:

	PA1	PA2
pre01	0.673	
pre02	0.926	
pre03	0.937	
pre04	0.934	
pre05	0.922	
pre06	0.665	
pro01		0.576
pro02		0.739
pro03		0.901
pro04		0.553
pro05	0.481	
pro06		0.772

	PA1	PA2
SS loadings	4.840	2.769
Proportion Var	0.403	0.231
Cumulative Var	0.403	0.634

¿Cuál es la interpretación de cada factor?

Una vez realizado el análisis factorial, se pueden obtener valores de los factores para cada observación

Para el ejemplo de foco regulatorio:

```
> pa.data.oblimin.scores <-  
+   factor.scores(focus, pa.data.oblimin, method="Thurstone")  
> scores <-  
+   as.data.frame(pa.data.oblimin.scores$scores)
```

Ejemplo: escalas de foco regulatorio

Puntuaciones factoriales

```
> colors <- rep("blue", 150)
> colors[which(data$sexo==1)] <- "red"
> plot(scores$PA1, scores$PA2, pch=19, col=colors, xlab="PA1 (prevention)", ylab="PA2 (promotion)",
+       cex.lab=0.7, cex.axis=0.7)
> abline(lm(scores$PA2 ~ scores$PA1))
```

