

Curso Estadística IV
Sociología
Universidad
Alberto Hurtado

Profesora
Carolina Aguilera
caguilera@uahurtado.cl



Ayudantes
Vicente Díaz – vidiazam@alumnos.uahurtado.cl
Miguel Tognarelli – mtognare@alumnos.uahurtado.cl



Clase 10

15 oct

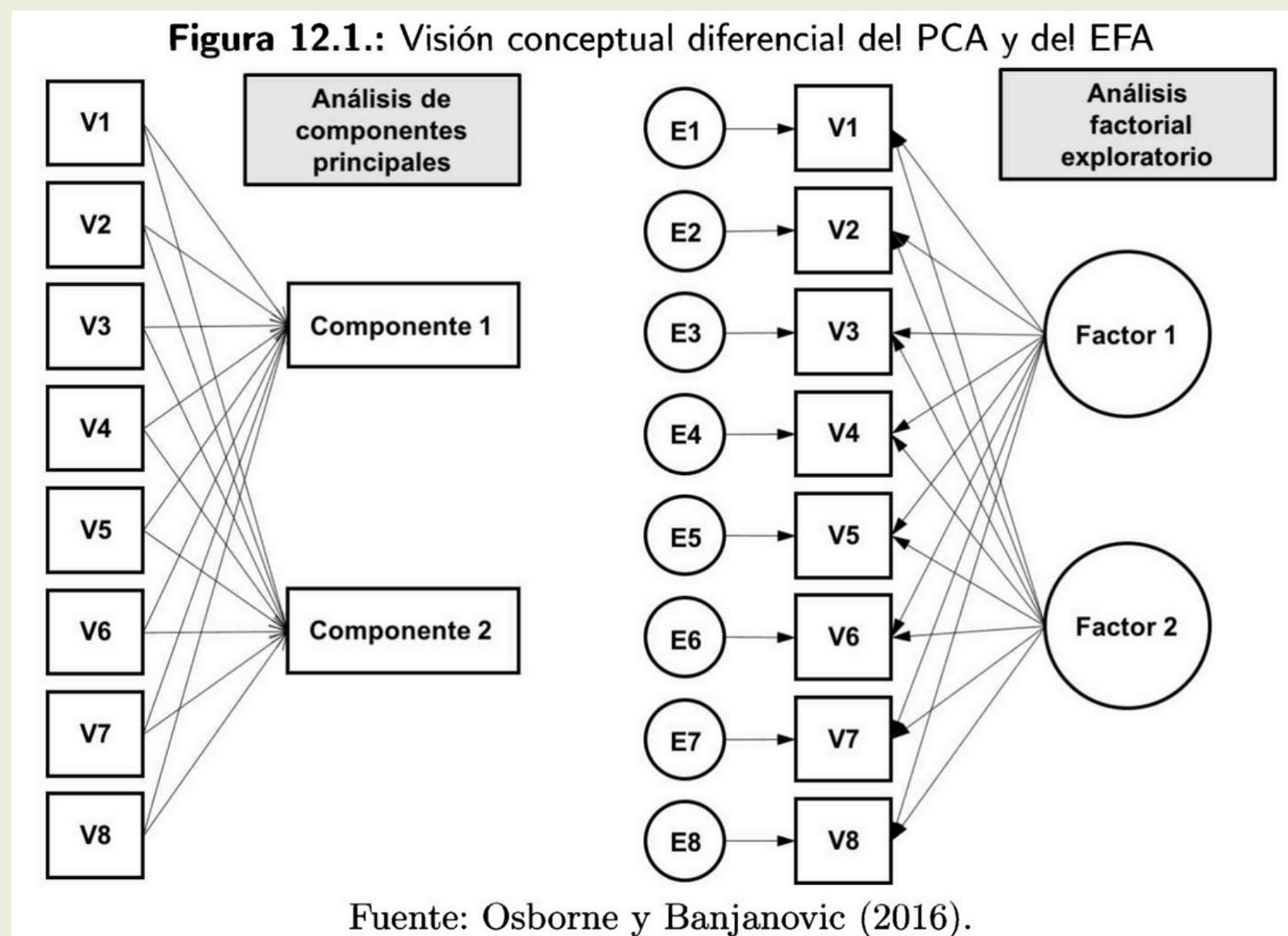
- Análisis Factorial exploratorio
- Ejemplo de aplicación
- Avanzando en el Trabajo Final



Repaso EFA

Considere un estudio simulado en que se analiza la cohesión barrial con 8 variables de escala de Likert que miden diferentes aspectos de relación con el barrio, con los siguientes resultados

barrio_ideal	Este es el barrio ideal para mí
Integrado	Me siento integrado/a en este barrio
Identifico	Me identifico con la gente de este barrio
parte_de_mi	Este barrio es parte de mí
amigos	En este barrio es fácil hacer amigos
sociable	La gente en este barrio es sociable
cordialidad	La gente en este barrio es cordial
colaboracion	La gente en este barrio es colaboradora



$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots \lambda_{1m}\xi_m + \varepsilon_1$$

$$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots \lambda_{2m}\xi_m + \varepsilon_2$$

...

$$x_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots \lambda_{pm}\xi_m + \varepsilon_p$$

χ_i = variables manifiestas (observadas)

λ_{ij} = es el peso del factor j en la variable i (causalidades)

ξ_i = son los factores comunes

ε_i = son factores únicos o errores

Ya no buscamos agrupar variables (PCA), sino descubrir la existencia de variables latentes que EXPLICAN la estructura latente de las variables. Se asume teóricamente que esto existe.

$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots \lambda_{1m}\xi_m + \varepsilon_1$$

$$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots \lambda_{2m}\xi_m + \varepsilon_2$$

...

$$x_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots \lambda_{pm}\xi_m + \varepsilon_p$$

χ^i = variables manifiestas (observadas)

λ^{ij} = es el peso del factor j en la variable i (cargas factoriales)

ξ^i = son los factores comunes

ε^i = son factores únicos o errores

Todas las variables originales vienen influidas por todos los factores comunes

Existe un factor único que es específico para cada variable.

Tanto los factores comunes como los factores únicos no son observables.

Fundamentos del EFA

Se estiman la o las variables latentes a un conjunto de indicadores, sin una especificación previa de la estructura factorial.

Preguntas a responder:

- ¿Cuántos factores subyacen a un conjunto de indicadores?
- ¿Cómo se relacionan los indicadores con los factores?
- ¿Cómo es la calidad del modelo estimado?

Fundamentos del EFA

Lo latente puede ser entendido como la varianza compartida por diferentes indicadores observados

La medición de variables latentes se encuentra asociada al modelo de factor común (Thurstone) y al análisis factorial

Pasos del EFA

- Estimación de matriz de correlaciones
- Pruebas medición de supuestos
- Decisión sobre número de factores
- Extracción de factores (varios métodos)
- Rotación
- Obtención de puntajes factoriales
- Interpretación y reporte

Ejemplo. Confianza en las Instituciones (ELSOC)

Considere un estudio simulado en que se analiza la cohesión barrial con 8 variables de escala de Likert que miden diferentes aspectos de relación con el barrio, con los siguientes resultados

barrio_ideal	Este es el barrio ideal para mí
Integrado	Me siento integrado/a en este barrio
Identifico	Me identifico con la gente de este barrio
parte_de_mi	Este barrio es parte de mí
amigos	En este barrio es fácil hacer amigos
sociable	La gente en este barrio es sociable
cordialidad	La gente en este barrio es cordial
colaboracion	La gente en este barrio es colaboradora

Matriz de Correlaciones							
barrio_ideal	integrado	identifico	parte_de_mi	amigos	sociable	cordialidad	colaboracion
1.00	0.70	0.63	0.67	0.37	0.36	0.45	0.35
0.70	1.00	0.74	0.72	0.47	0.45	0.49	0.42
0.63	0.74	1.00	0.73	0.49	0.48	0.52	0.44
0.67	0.72	0.73	1.00	0.45	0.43	0.48	0.41
0.37	0.47	0.49	0.45	1.00	0.65	0.54	0.48
0.36	0.45	0.48	0.43	0.65	1.00	0.67	0.56
0.45	0.49	0.52	0.48	0.54	0.67	1.00	0.57
0.35	0.42	0.44	0.41	0.48	0.56	0.57	1.00

Interpretación de R (matriz de correlaciones).

Valores de correlación cercanos a 1 = existe correlación, valores cercanos a 0 = no existe correlación

- identifico_m con integrado_m = 0.58: Personas que se identifican con el barrio consideran sentirse integradas.
- sociable_m con cordialidad = 0.67 Quienes perciben más facilidad para hacer amigos tienden a considerar a la gente del barrio cordial.

Entonces, tiene sentido hacer un EFA, para buscar variables latentes (factores).

Paso 2. Supuesto de adecuación de matriz para AFE KMO (Kaiser, Meyer, Olkin Measure of Sampling Adequacy):

1. Test de adecuación de matriz para AFE

Varía entre 0 y 1. Contrastá si las correlaciones parciales entre las variables son pequeñas

Valores pequeños (menores a 0.5) indican que los datos no serían adecuados para AFE, ya que las correlaciones entre pares de variables no pueden ser explicadas por otras variables.

```
corMat <- proc_data %>% select(barrio_ideal:colaboracion) %>%  
  cor(use = "complete.obs") # estimar matriz pearson  
  
KMO(corMat)
```

```
## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = corMat)  
## Overall MSA =  0.9  
## MSA for each item =  
## barrio_ideal    integrado   identifico  parte_de_mi      amigos    sociable  
##        0.90        0.89       0.90       0.90       0.90       0.85  
##  cordialidad colaboracion  
##        0.90        0.93
```

Interpretación: el valor Overall MSA = 0,9, muy cercano a 1, así que las variables si pueden explicarse por variables latentes. Ninguna variable individual arroja un valor de 0,5 o menos, por lo que se pueden considerar todas las variables

Varía entre 0 y 1. Contrastá si las correlaciones parciales entre las variables son pequeñas

Valores pequeños (menores a 0.5) indican que los datos no serían adecuados para AFE, ya que las correlaciones entre pares de variables no pueden ser explicadas por otras variables.

Paso 2. Supuesto nivel de correlaciones de la matriz: test de esfericidad de Barlett

```
cortest.bartlett(corMat, n = 3417)
```

```
## $chisq  
## [1] 16081.04  
##  
## $p.value  
## [1] 0  
##  
## $df  
## [1] 28
```

Interpretación: el valor $p = 0$, por lo que se rechaza la H_0 , y se concluye que sí hay significación estadística de las correlaciones

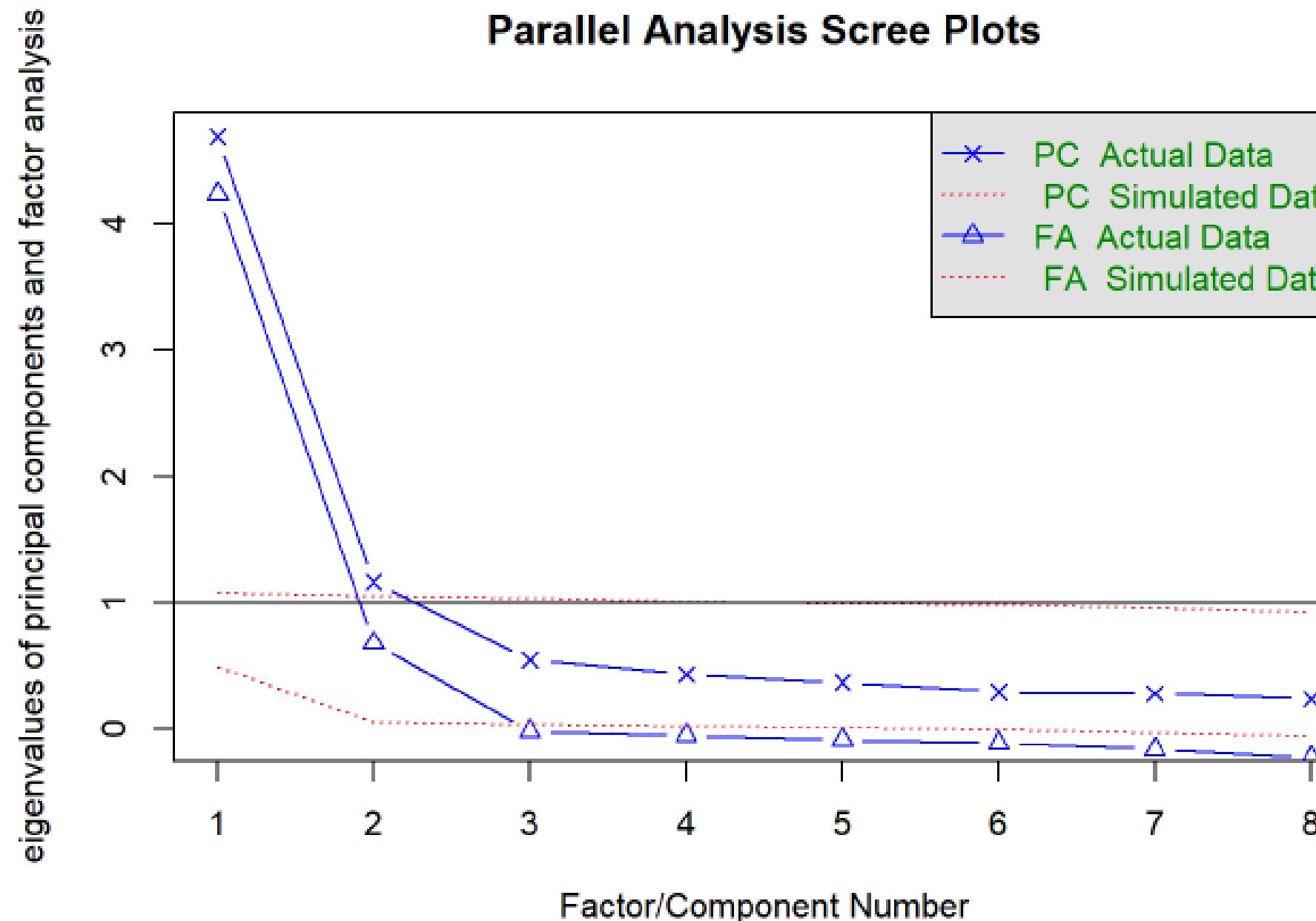
Se utiliza para evaluar la hipótesis que la matriz de correlaciones es una matriz identidad (diagonal=1 y bajo la diagonal=0): las variables no se correlacionan entre sí.

Se busca significación ($p < 0.05$) que indica que las variables estén correlacionadas

Paso 3. Definición de número de factores. Método de Análisis Paralelo

Análisis paralelo para decidir número de factores:

```
fa.parallel(corMat, n.obs=3417)
```



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

Interpretamos: el número de factores a seleccionar es 2. Este análisis es más robusto y nos quedamos con este resultado.

¿Qué nos dice este método?

¿Con cuántos factores nos quedamos?

Paso 4. Extracción de los factores con el método de Maximum Likelihood. Considerando que se considerarán 2 factores

Método Maximum likelihood

Maximiza la posibilidad de que los parámetros reproduzcan los datos observados

```
fac_ml <- proc_data %>% select(barrow_ideal:colaboracion) %>% fa(nfactors = 2, fm= "ml")  
fac_ml
```

```
## Factor Analysis using method = ml  
## Call: fa(r = ., nfactors = 2, fm = "ml")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
##          ML1    ML2    h2   u2 com  
## barrio_ideal  0.82 -0.06  0.61  0.39  1.0  
## integrado     0.86  0.01  0.76  0.24  1.0  
## identifico    0.79  0.09  0.72  0.28  1.0  
## parte de mi   0.85  0.02  0.71  0.29  1.0  
## amigos        0.10  0.67  0.54  0.46  1.0  
## sociable      -0.09  0.94  0.78  0.22  1.0  
## cordialidad   0.12  0.69  0.61  0.39  1.1  
## colaboracion  0.10  0.60  0.44  0.56  1.1  
  
##  
##          ML1    ML2  
## SS loadings  2.90  2.28 Suma de cuadrados de las cargas factoriales → mide la “fuerza” de cada factor.  
## Proportion Var 0.36  0.28 Porcentaje de la varianza total explicada por cada factor.  
## Cumulative Var 0.36  0.65 Porcentaje acumulado de varianza explicada hasta ese factor.  
## Proportion Explained 0.56  0.44 Proporción de la varianza explicada solo considerando los factores retenidos.  
## Cumulative Proportion 0.56  1.00 Porcentaje acumulado (siempre llega a 1 = 100%).  
##
```

Tenemos dos factores que explican la estructura de los datos. En este caso, a diferencia del PCA, la variable barrio_ideal está correlacionada fuertemente con el ML1.

Ambos factores tienen una fuerza similar, aunque el ML1 es mayor. El % de la varianza total explicada por MLA es 36%, y si se toman ambas como el total el ML1 explica el 56% de la varianza total.

Paso 4. Extracción de los factores con el método de Maximum Likelihood. Considerando que se considerarán 2 factores

Método Maximum likelihood

Maximiza la posibilidad de que los parámetros reproduzcan los datos observados

```
fac_ml <- proc_data %>% select(barrido_ideal:colaboracion) %>% fa(nfactors = 2, fm= "ml")
fac_ml
```

```
## Factor Analysis using method = ml
## Call: fa(r = ., nfactors = 2, fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          ML1    ML2    h2   u2 com
## barrio_ideal  0.82 -0.06  0.61  0.39 1.0
## integrado     0.86  0.01  0.76  0.24 1.0
## identifico    0.79  0.09  0.72  0.28 1.0
## parte_de_mi    0.85  0.00  0.71  0.29 1.0
## amigos         0.10  0.67  0.54  0.46 1.0
## sociable      -0.09  0.94  0.78  0.22 1.0
## cordialidad    0.12  0.69  0.61  0.39 1.1
## colaboracion   0.10  0.60  0.44  0.56 1.1
##
##          ML1    ML2
## SS loadings   2.90  2.28
## Proportion Var 0.36  0.28
## Cumulative Var 0.36  0.65
## Proportion Explained 0.56  0.44
## Cumulative Proportion 0.56 1.00
##
```

h2 (comunalidad): proporción de varianza explicada por todos los factores juntos (ideal > 0.4, cercano a 1).

u2 (unicidad): parte no explicada (1 - h2).

com (complejidad): cuántos factores influyen en una variable (1 = carga simple; > 2 = más compleja).

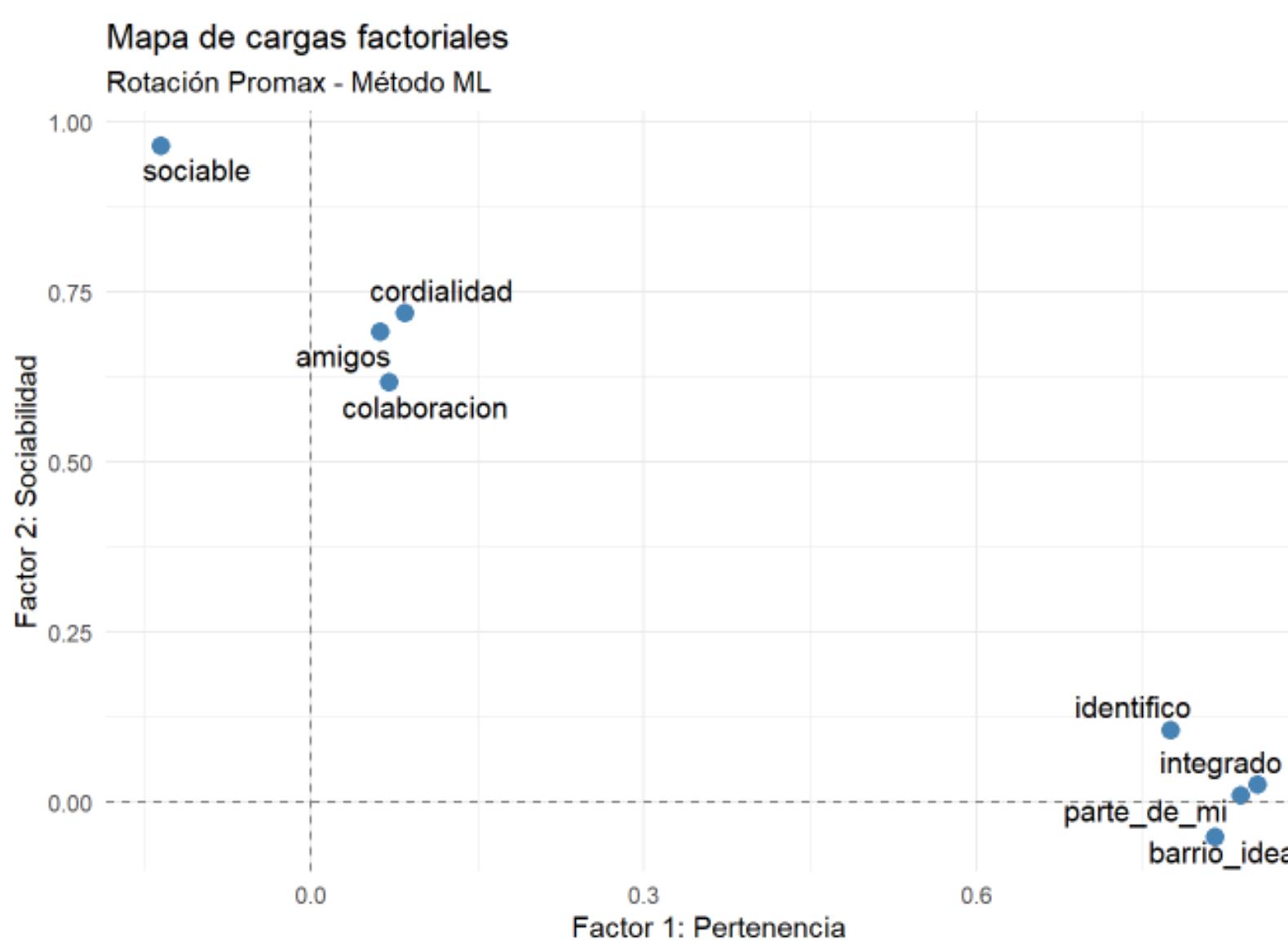
Paso 4. Extracción de los factores con el método de Maximum Likelihood.

Considerando que se considerarán 2 factores

```
# With factor correlations of  
#          ML1   ML2  
# ML1  1.00  0.64  
# ML2  0.64  1.00
```

Matriz de correlación entre factores. Justifica rotación oblicua

Comparamos EFA con



	ML1	ML2
SS loadings	2.90	2.28
Proportion Var	0.36	0.28
Cumulative Var	0.36	0.65
Proportion Explained	0.56	0.44
Cumulative Proportion	0.56	1.00

variable	ML1	ML2
barrio_ideal	0,815085	-0,05112
integrado	0,8539	0,02611
identifico	0,774921	0,107285
parte_de_mi	0,838834	0,010005
amigos	0,062257	0,693055
sociable	-0,13582	0,965324
cordialidad	0,084577	0,719777
colaboracion	0,069957	0,617681

Paso 6. Puntajes factoriales

Una vez que tenemos el modelo factorial, con las cargas factoriales, podemos calcular para cada persona (o caso) un puntaje que aproxima su valor en cada factor latente.

Combinar con los datos originales

```
proc_data_scores <- cbind(proc_data, fac_ml$scores)
```

Vemos los valores para los primeros 10 casos:

```
head(proc_data_scores[, c("pertenencia", "sociabilidad")], 10)
```

```
##      pertenencia sociabilidad
## 1   -0.19031088    0.4035489
## 2   -0.03008579    0.4594585
## 3   -0.88757110   -0.6815509
## 4    0.29062166   -0.3395309
## 5   -0.17697314   -0.3497488
## 6    0.33641159    0.4762188
## 7    0.33641159    0.4762188
## 8    0.33641159    0.4762188
## 9    0.30830384    0.3108405
## 10  -0.51783704   -0.1316368
```

Ejercicio 1

(0,4 puntos para Control 2)

Explique las razones y procedimiento de esta construcción de un índice de cohesion social

<https://ocscoes.github.io/medicion-cohesion-LA/intro.html>

Ejercicio Módulo 2

Propone una investigación sociológica que pueda ser respondida por (al menos) Análisis de Correspondencias y/o Análisis de Componentes principales o por Análisis Factorial Exploratorio.

(0,2 puntos Control 2)

Incluya:

Pregunta de investigación

Dimensiones del análisis y posibles variables

Base de datos y módulos que permitieran medir esas variables

Propuesta de Modelo justificado