Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 10: 23/09/24 – 28/09/24

Motivación

Existen **variables que no se pueden medir, definir u observar**, las cuales son llamadas variables latentes; sin embargo, existen propuestas para intentar medirla.

Motivación

Existen **variables que no se pueden medir, definir u observar**, las cuales son llamadas variables latentes; sin embargo, existen propuestas para intentar medirla. La **inteligencia** es un ejemplo claro de variables latentes y suele "medirse" con entrevistas, cuestionarios, exámenes, entre otros. Por ejemplo:

- El **ICFES** intenta comprender la inteligencia de los estudiantes de básica media en algunas áreas como matemáticas, lenguaje, ciencias, inglés, etc.
- En el **proceso de selección** de una empresa, se tienen pocas vacantes y muchos candidatos. La selección se hace aplicando cuestionarios que envuelven pruebas objetivas (específicas para el cargo), entrevistas, etc.

Motivación

Aunque el conjunto de pruebas no definen completamente la inteligencia de la persona, sabemos que las personas con mayor conocimiento suelen tener mejor desempeño en las pruebas.

La **clase social** y **el nivel de desarrollo de un país** son otros dos ejemplos de **variables latentes**. En el primer caso, el sociólogo podría plantear preguntas sobre la ocupación o cargo de la persona, el nivel educativo, si tiene vivienda, etc, suponiendo que éstas reflejan el concepto que realmente le interesa (clase social). En el segundo, se pueden terner encuenta el PIB del país, acceso a la educación, etc.

En general, las **variables latentes suelen encontrarse en varias áreas** de la psicología, educación, economía, química, marketing, geología, entre otras.

Introducción

Los investigadores recopilan información sobre las **variables (mensurables o manifiestas)** que a su juicio considera **relevantes** o hacen visibles los conceptos de interés (**variables latentes**). Puntualmente, el investigador trata de descrubir si las relaciones entre las variables observadas son consistentes con lo que se quiere que ellas midan.

Introducción

Los investigadores recopilan información sobre las **variables (mensurables o manifiestas)** que a su juicio considera **relevantes** o hacen visibles los conceptos de interés (**variables latentes**). Puntualmente, el investigador trata de descrubir si las relaciones entre las variables observadas son consistentes con lo que se quiere que ellas midan. Entonces, para resolver el problema, **la idea básica** fue **agrupar las variables de acuerdo con sus correlaciones**. Específicamente, se asume o se "controla" que todas las variables dentro de un grupo particular están altamente correlacionadas entre sí y tienen correlaciones relativamente bajas con variables de otro grupo diferente.

Introducción

Charles Spearman y Karl Pearson en 1904 propusieron y usaron esa idea para intentar definir y medir la **inteligencia** y obtuvieron dos grupos:

- El primero conformado por las **puntuaciones de pruebas de clásicos, francés, inglés, matemáticas y música**, las cuales estaban **altamente correlacionadas**.
- El segundo conformado por las variables que representaban las **puntuaciones de aptitud**.

El primer grupo puede ser considerado como un factor de inteligencia y el segundo como factor de aptitud.

Introducción

El método de análisis más utilizado para ayudar a descubrir las relaciones entre las variables latentes supuestas y las variables manifiestas es el **Análisis Factorial (AF)**.

Objetivos de Análisis Factorial

- ✓ El objetivo principal del AF es **describir la variabilidad original del vector aleatorio $\mathbf{X}_{p\times 1}$, en términos de** un número $\underline{m}$ de variables aleatorias llamadas **factores comunes** y que están relacionadas con $\mathbf{X}_{p\times 1}$ a través de un modelo lineal[1]. En el modelo, parte de la **variabilidad de $\mathbf{X}_{p\times 1}$** es atribuida a los **factores comunes** y el **resto de la variabilidad de $\mathbf{X}_{p\times 1}$** se atribuye a las variables que no fueron incluidas en el modelo, es decir, el **error aleatorio**.
- ✓ **Agrupar las variables de $\mathbf{X}_{p\times 1}$ en factores mutuamente no correlacionados, interpretables**. Encontrar estos factores de agrupamiento es uno de los objetivos del AF.

[1] las variables observadas se hacen en regresión sobre los factores.

Objetivos de Análisis Factorial

Observaciones:

1. *Evidentemente, la idea es* ***determinar*** *un número m* ***pequeño de factores, no correlacionados*** *que de algún modo resuman las informaciones principales de las variables originales.*
2. *Es importante destacar que* ***dentro del AF*** *tenemos dos tipos de análisis:*
   - ✓ ***Exploratorio****: se buscan los factores subyacentes a las variables originales.*
   - ✓ ***Confirmatorio****: se tiene un modelo factorial (hipotético) y se quiere verificar si es consistente con los datos disponibles.*

Análisis Factorial

El primer paso **para aplicar el AF es comprobar si su uso es válido para nuestros datos**. Para ello podemos utilizar dos métodos:

1. La **Pueba de esfericidad de Bartlett:** se prueba la hipótesis nula de que las variables no están correlacionadas:

$$H_0 : \mathbf{P} = \mathbf{I}$$
$$H_1 : \mathbf{P} \neq \mathbf{I}$$

Evidentemente el uso o **aplicación del AF es válido cuando** H_0 **es rechazada**.

```
# ----- Prueba de esfericidad de Bartlett ----- #
require(psych)
cortest.bartlett(R)
```

Análisis Factorial

- El **criterio de Kaiser-Meyer-Olkin (KMO):** compara las magnitudes de los coeficientes de correlación observados con las magnitudes de los coeficientes de correlación parcial:

$$\text{KMO} = \frac{\sum_{i \neq j}^{p} r_{ij}^2}{\sum_{i \neq j}^{p} r_{ij}^2 + \sum_{i \neq j}^{p} a_{ij}^2}$$

donde r_{ij} es la correlación muestral entre X_i y X_j, y a_{ij} es la correlación parcial entre X_i y X_j. **Valores pequeño de** KMO **indica que el AF no es apropiado**.

Análisis Factorial

En particular, los autores hacen las siguientes **recomendaciones**:

KMO	**Evaluación**
≥ 0.9	Excelente
$[0.8, 0.9)$	Buena
$[0.7, 0.8)$	media
$[0.6, 0.7)$	Baja
$[0.5, 0.6)$	Mala
< 0.5	inaceptable

Puntualmente, el uso o aplicación del AF es válido cuando KMO ≥ 0.7.

```
# ----- Criterio KMO ----- #
require(psych)
 KMO(R)
```

Análisis Factorial

Example

Se aplicó una prueba psicológica en $n = 145$ niños de básica primaria. En el cuestionario comprendía la comprensión de: párrafos (PARA), oraciones/frases (SENT), significado de palabras (WORD), sumar (ADD), contar (DOTS). De acuerdo con los resultados se obtuvo la **matriz de correlación**

	PARA	SENT	WORD	ADD	DOTS
PARA	1				
SENT	0.722	1			
WORD	0.714	0.685	1		
ADD	0.203	0.246	0.170	1	
DOTS	0.095	0.181	0.113	0.585	1

Análisis Factorial

Observe que:

- Las variables SENT, PARA y Word tienen **correlaciones relativamente altas y referentes a lenguaje**.
- Las varables ADD y DOTS tienen una **correlación moderada y relativas a matemáticas**.
- Las **variables de lenguaje tienen baja correlación con las variables de matemáticas**.

Entonces, lo que se quiere con estas variables es tener un factor (podría ser inteligencia) que explique cada una de esas variables.

Análisis Factorial

Denotando cada una de las variables por X_i, $i = 1, \ldots, 5$, se estaría pensando que **cada variable va ser una función** de un factor (F) y una variación aleatoria (ϵ_i):

$$X_i = g(\mathrm{F}, \epsilon_i) = l_i \mathrm{F} + \epsilon_i \qquad i = 1, 2, \ldots, 5 \tag{1}$$

esto es,

$$\begin{aligned}
\text{PARA} &\longrightarrow X_1 = l_1 \mathrm{F} + \epsilon_1 \\
\text{SENT} &\longrightarrow X_2 = l_2 \mathrm{F} + \epsilon_2 \\
\text{WORD} &\longrightarrow X_3 = l_3 \mathrm{F} + \epsilon_3 \\
\text{ADD} &\longrightarrow X_4 = l_4 \mathrm{F} + \epsilon_4 \\
\text{DOTS} &\longrightarrow X_5 = l_5 \mathrm{F} + \epsilon_5
\end{aligned}$$

Análisis Factorial

La ecuación (1) es un **modelo de regresión** con

- **variable respuesta** X_i.
- l_i es el efecto/importancia (**carga factorial**) del factor F para la $i-$ésima variable observada.
- ϵ_i una variación aleatoria.

Si F es el factor inteligencia, éste intervendrá en el efecto de cada una de las variables. Por ejemplo, a **mayor inteligencia**, **mayor será la habilidad** de la persona en esos conceptos.

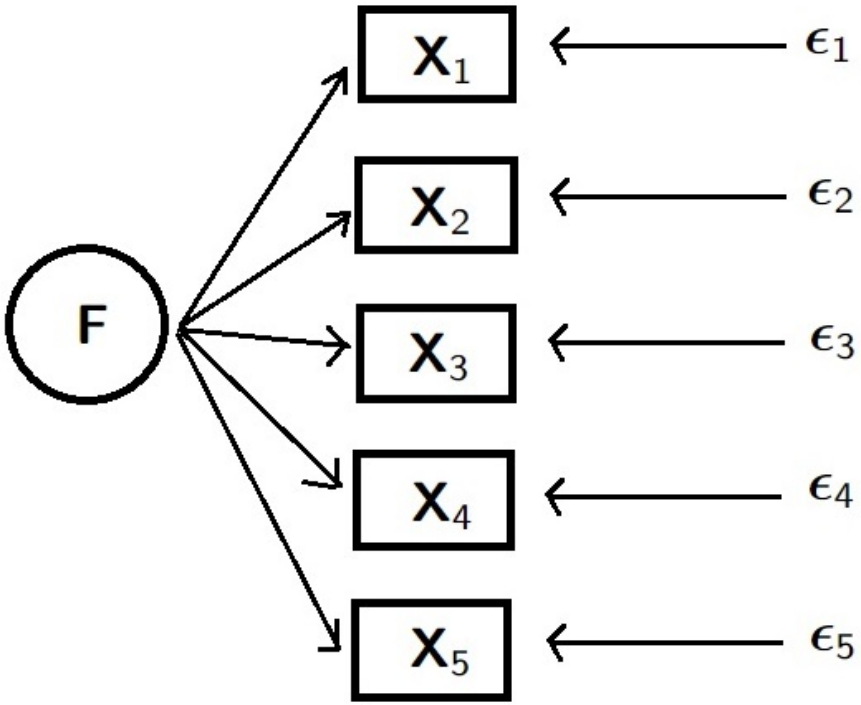
Análisis Factorial

La ecuación (1) es un **modelo de regresión** con

- **variable respuesta** X_i.
- l_i es el efecto/importancia (**carga factorial**) del factor F para la $i-$ésima variable observada.
- ϵ_i una variación aleatoria.

Si F es el factor inteligencia, éste intervendrá en el efecto de cada una de las variables. Por ejemplo, a **mayor inteligencia**, **mayor será la habilidad** de la persona en esos conceptos. Por lo tanto, se puede explicar una variación única/común (en todas las 5 variables) que se debe a F y otra específica de cada variables (ϵ_i).

Análisis Factorial



Observe que F afecta a $X_1, \ldots, X_5$ y que además de ese factor, tenemos otra variable, $\epsilon_1, \ldots, \epsilon_5$ que es una variación específica de cada una de las X_i's.

Análisis Factorial

Considere que se va trabajar con las variables X_i **estandarizadas**:

$$Z_i = g(\mathsf{F}, \epsilon_i) = l_i \mathsf{F} + \epsilon_i \qquad i = 1, 2, \ldots, 5$$

Además, considere que $\text{Var}[\mathsf{F}] = 1$, ϵ_i y F son **no correlacionados**. Luego

$$\begin{aligned} 1 &= \text{Var}[Z_i] \\ &= \text{Var}[l_i \mathsf{F} + \epsilon_i] \\ &= l_i^2 \, \text{Var}[\mathsf{F}] + \text{Var}[\epsilon_i] \\ &= l_i^2 + \psi_i \end{aligned}$$

Análisis Factorial

así,

- l_i^2 es la proporción de variación de Z_i explicada por F (factor común)
- ψ es la varianza restante (específica) de Z_i.

Generalizando!

Considere el vector aleatorio

$$\mathbf{X}_{p\times 1} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

Análisis Factorial

con

- Vector de medias $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}_{p\times 1}$
- Matriz de covarianzas $\mathrm{Var}[\mathbf{X}] = \boldsymbol{\Sigma}_{p\times p}$
- Matriz de correlación $\mathrm{Cor}[\mathbf{X}] = \mathbf{P}_{p\times p}$

y considere las estandarizaciones de las X_i

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \qquad i = 1, \ldots, p$$

Sabemos que

$$\mathrm{Cov}[\mathbf{Z}] = \mathbf{P}_{p\times p} = \mathrm{Corr}[\mathbf{X}]$$

Análisis Factorial

Queremos explicar $Z_1, \ldots, Z_p$ con $F_1, \ldots, F_m$ factores de tal forma que

$$Z_1 = l_{11}F_1 + \cdots + l_{1m}F_m + \epsilon_1$$

$$Z_2 = l_{21}F_1 + \cdots + l_{2m}F_m + \epsilon_2$$

$$\vdots$$

$$Z_p = l_{p1}F_1 + \cdots + l_{pm}F_m + \epsilon_p$$

Análisis Factorial

Matricialmente,

$$\underbrace{\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}}_{\mathbf{Z}_{p\times 1}} = \underbrace{\begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix}}_{\mathbf{L}_{p\times m}} \underbrace{\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}}_{\mathbf{F}_{m\times 1}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}}_{\boldsymbol{\epsilon}_{p\times 1}}$$

esto es,

$$\mathbf{Z}_{p\times 1} = \mathbf{L}_{p\times m}\ \mathbf{F}_{m\times 1} + \boldsymbol{\epsilon}_{p\times 1} \tag{2}$$

Análisis Factorial

donde

- **L** es la matriz de **cargas factoriales**.
- l_{ij} son los **elementos de L**, llamados loading (pesos) o carga factorial, y corresponde al coeficiente de Z_i en F_j, $i = 1, \ldots, p$ y $j = 1, \ldots, m$.
- **F** contiene los m factores (no observables o latentes) que describen los elementos de la población.
- ϵ es el vector de errores aleatorios (no observables).

Análisis Factorial

Supuestos del modelo:

- $\mathrm{E}[\mathbf{F}] = \mathbf{0}$
- $\mathrm{Var}[\mathbf{F}] = \mathbf{I}$
- $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$
- $\mathrm{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$
- $\mathrm{Cov}[\mathbf{F}, \boldsymbol{\epsilon}] = \mathbf{0}$

Como los *m* factores son ortogonales entre sí, el modelo (2) es llamado **Modelo Factorial Ortogonal**.

Análisis Factorial

El objetivo es **identificar** las nuevas m **variables** $(\mathrm{F}_j,\, j = 1, \ldots, m)$, **interpretarlas** y calcular sus **scores**, de forma análoga como se hizo en ACP.

Recuerde que el interés es **explicar la variabilidad** de

$$\mathbf{Z}_{p\times 1} = \mathbf{L}_{p\times m}\ \mathbf{F}_{m\times 1} + \boldsymbol{\epsilon}_{p\times 1}$$

y que

$$\mathrm{Cov}[\mathbf{Z}] = \mathbf{P} = \mathrm{Cor}[\mathbf{X}]$$

Luego

$$\mathbf{P} = \mathrm{Cov}[\mathbf{Z}] = \mathrm{Cov}\left[\mathbf{L}\ \mathbf{F} + \boldsymbol{\epsilon}\right]$$

Análisis Factorial

$$\mathbf{P} = \mathbf{L}\,\text{Cov}[\mathbf{F}]\,\mathbf{L}^\top + \text{Cov}[\boldsymbol{\epsilon}]$$

$$= \mathbf{L}\,\mathbf{I}\,\mathbf{L}^\top + \boldsymbol{\Psi}$$

$$= \mathbf{L}\,\mathbf{L}^\top + \boldsymbol{\Psi}$$

Así que el objetivo es encontrar $\mathbf{L}$ y $\boldsymbol{\Psi}$ de tal forma que $\mathbf{P}$ pueda descomponerse como $\mathbf{L}\,\mathbf{L}^\top + \boldsymbol{\Psi}$, lo cual no siempre es posible.

Análisis Factorial

Note que:

$$\mathbf{P} = \mathbf{L}\,\mathbf{L}^{\top} + \mathbf{\Psi}$$

$$\begin{bmatrix} 1 & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{1p} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} l_{1j}^2 & \cdots & \sum_{j=1}^{m} l_{1j} l_{jp} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{m} l_{pj} l_{j1} & \cdots & \sum_{j=1}^{m} l_{pj}^2 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

Análisis Factorial

de ahí

- ✓ $1 = \text{Var}[Z_i] = \sum_{j=1}^{m} l_{ij}^2 + \psi_i = h_i^2 + \psi_i$, donde
  - h_i^2 es llamado comunalidad (por los factores), o sea, cuanto los factores (que se estimaron) logran explicar de la varianza de Z_i.
  - ψ_i es la varianza específica, o sea, lo que los factores no lograron explicar de la variable.
- ✓ $\text{Cov}[Z_i, Z_k] = l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{im}l_{km}$, $i, k = 1, 2, \ldots, p$ $i \neq k$.

Análisis Factorial

- ✓ $\text{Cov}[\mathbf{Z}, \mathbf{F}] = \mathbf{L} \longrightarrow \text{Cov}[Z_i, F_j] = \text{Cor}[Z_i, F_j] = l_{ij}$, $i = 1, \ldots, p$ y $j = 1, \ldots, m$.
  - **L** puede ser utilizada en la búsqueda del entendimiento e interpretación de los factores.
  - **En la práctica**, para realizar AF, lo **primero** que se debe hacer es utilizar mecanismos adecuados para **estimar** m, para **posteriormente estimar L** y $\boldsymbol{\Psi}$.
- ✓ La **varianza total explicada por el factor** $\mathbf{F}_j$ está dada por

$$\text{PVTE}_{F_j} = \frac{\sum_{i=1}^{p} l_{ij}^2}{p}$$

Análisis Factorial

Existen varios **criterios para escoger** $\underline{m}$, el número de factores:

1. Análisis de la **proporción de varianza total** relacionada con cada uno de los autovalores λ_i asociados a la matriz de correlación **R**, que está dada por

$$\frac{\lambda_i}{p} \qquad i = 1, 2, \ldots, p.$$

2. Número de autovectores de la matriz de correlación muestral **R** que son mayores que 1 (criterio de **Kaiser**).
3. Gráfico de codo (**scree plot**).

Análisis Factorial

Observación: Los tres criterios llevan en cuenta la naturaleza (grandeza) numérica de los autovalores. Sin embargo, una ***selección adecuada****, también debe* ***incluir*** *la* ***interpretación de los factores*** *y el* ***principio de parsimonia****.*

Análisis Factorial

Observación: Los tres criterios llevan en cuenta la naturaleza (grandeza) numérica de los autovalores. Sin embargo, una ***selección adecuada****, también debe* ***incluir*** *la* ***interpretación de los factores*** *y el* ***principio de parsimonia****.*

Una vez definido m, $\mathbf{L}$ y $\mathbf{\Psi}$ **pueden ser estimados por**:

1. Método de los **factores** (o componentes) **principales** (ACP en término de los factores)
2. Método de los **factores** (o componentes) **principales iterativo**.
3. Método de **máxima verosimilitud** (asumiendo normalidad)

1. Método de los Factores Principales

Consiste en hacer una descomposición espectral de $\mathbf{P}$.

Si $m = p$, sigue que

$$\mathbf{P} = \sum_{i=1}^{p} \lambda_i \, \mathbf{e}_i \, \mathbf{e}_i^\top$$

$$= \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1 & \dots & \sqrt{\lambda_p}\,\mathbf{e}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1 \\ \vdots \\ \sqrt{\lambda_p}\,\mathbf{e}_p \end{bmatrix}$$

$$= \mathbf{L}^\top \mathbf{L}$$

Análisis Factorial

donde λ_i son los autovalores de $\mathbf{P}$ (o $\mathbf{R}$) y $\mathbf{e}_i$ el autovector asociado al correspondiente autovalor.

En la práctica el interés es $m < p$

En ese caso, tenemos el siguiente algoritmo:

- Extraer λ_i y $\mathbf{e}_i$ de $\mathbf{R}$

$$(\lambda_i, \mathbf{e}_i) \qquad i = 1, 2, \ldots, p$$

- Selecionar los m autovalores y autovectores correspondientes

$$(\lambda_i, \mathbf{e}_i) \qquad i = 1, 2, \ldots, m$$

Análisis Factorial

- Estimar $\mathbf{L}$ y $\boldsymbol{\Psi}$:

$$\hat{\mathbf{L}} = \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \quad \dots \quad \sqrt{\lambda_m}\,\mathbf{e}_m\right]$$

$$\hat{\boldsymbol{\Psi}} = \text{Diag}\left(\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}^\top\right)$$

Observación: $MRES = \boldsymbol{R} - \left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^\top + \hat{\boldsymbol{\Psi}}\right)$ *es la* ***matriz residual***.

Bajo éste método, la **proporción de varianza explicada** por F_j queda dada por:

$$\text{PVE}_{\mathsf{F}_j} = \frac{\lambda_i}{p}$$

Análisis Factorial

Example

Se llevó a una pesquisa de mercado para evaluar la **satisfacción de un consumidor** con relación a un nuevo producto. El cuestionario se aplicó en **200 personas** y contenía **preguntas sobre el sabor, aroma, color, textura, utilidad, facilidad de encontrar y empaque**. Para estos datos, se obtuvo la **matriz de correlación**:

Análisis Factorial

Example

	Sabor	Aroma	Color	Textura	Utilidad	Local	Empaque
Sabor	1						
Aroma	0.969	1					
Color	0.801	0.711	1				
Textura	0.960	0.991	0.696	1			
Utilidad	0.050	0.037	0.138	0.029	1		
Local	0.056	0.046	0.096	0.045	0.833	1	
Empaque	0.103	0.090	0.163	0.087	0.693	0.530	1

Realice un análisis factorial.

Análisis Factorial

Solución:

Antes de realizar el análisis factorial note que a partir de la matriz de correlación:

- Las variables sabor, aroma, color y textura estan altamente correlacionadas y poco correlacionadas con las demás.
- Las variables utilidad, local y empaque estan correlacionadas y poco correlacionadas con el resto.

Ahora verificamos si el análisis factorial es viable, para ello, verificaremos si las variables no están correlacionadas:

$$H_0 : \mathbf{P} = \mathbf{I}$$
$$H_1 : \mathbf{P} \neq \mathbf{I}$$

Análisis Factorial

```
# ----- Matriz de Correlación Muestral ----- #
R<-matrix(c(1,0.969,0.801,0.96,0.05,0.056,0.103,
            0.969,1,0.711,0.991,0.037,0.046,0.09,
            0.801,0.711,1,0.696,0.138,0.096,0.163,
            0.96,0.991,0.696,1,0.029,0.045,0.087,
            0.05,0.037,0.138,0.029,1,0.833,0.693,
            0.056,0.046,0.096,0.045,0.833,1,0.53,
            0.103,0.09,0.163,0.087,0.693,0.53,1),
          ncol=7)

# --- Prueba de esfericidad de Bartlett
require(psych)
cortest.bartlett(R)
```

Análisis Factorial

```
$chisq
[1] 956.883

$p.value
[1] 4.248392e-189

$df
[1] 21
```

observe que

$$p - \text{valor} = 4.248392e - 189 < 0.05 = \alpha$$

así que **hay suficiente evidencia** en la muestra para rechazar H_0, en consecuencia, **existe correlación entre las variables**, por lo tanto el **AF es viable**.

Análisis Factorial

El **criterio KMO**

```
# --- KMO
KMO(R)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = R)
Overall MSA =  0.72
MSA for each item =
[1] 0.80 0.72 0.78 0.76 0.57 0.61 0.73
```

También nos indica que el AF es viable (KMO $= 0.72$).

Análisis Factorial

Para e AF, primero debemos **determinar el número de factores** m. Sabemos que bajo el método de los **factores principales**, la proporción de varianza explicada por F_j está dada por:

$$\text{PVE}_{F_j} = \frac{\lambda_i}{p}$$

```
auto<-eigen(R)
lambdai<-auto$values  # --- autovalores

PVE<-lambda/7                  # prop. de var. expl.
round(PVE,4)
[1] 0.5191 0.3332 0.0695 0.0540 0.0190 0.0039 0.0012
```

Análisis Factorial

```
PVE_acum<-cumsum(PVE)     # prop. de var. expl. acum.
round(PVE_acum,4)
[1] 0.5191 0.8523 0.9219 0.9759 0.9949 0.9988 1.0000
```

Observe que

- el primer factor (F_1), explica el 51.91 % de la varianza total,
- el segundo (F_2), explica el 33.32 %, de la varianza total
- F_1 y F_2 explican el 85 % de la varianza total

Por lo tanto, podemos escoger $m = 2$ **factores.**

Análisis Factorial

Luego, la **matriz de cargas factoriales** está dada por

$$\hat{\mathbf{L}} = \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1 & \sqrt{\lambda_2}\,\mathbf{e}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{\lambda_1}\begin{bmatrix} e_{11} \\ \vdots \\ e_{17} \end{bmatrix} & \sqrt{\lambda_2}\begin{bmatrix} e_{21} \\ \vdots \\ e_{27} \end{bmatrix} \end{bmatrix}$$

```
e<-auto$vectors     # autovectores

# --- Matriz de Cargas Factoriales
L<-e*matrix(rep(sqrt(lambdai),7),ncol = 7,byrow = T)
```

Análisis Factorial

```
L_m<-L[,1:2]*(-1) # Escogemos los m=2 primeros factores

round(L_m,4)
       [,1]    [,2]
[1,] 0.9738 -0.1744
[2,] 0.9587 -0.1885
[3,] 0.8410 -0.0513
[4,] 0.9521 -0.1920
[5,] 0.2175  0.9282
[6,] 0.2040  0.8676
[7,] 0.2526  0.7828
```

Análisis Factorial

Por lo tanto

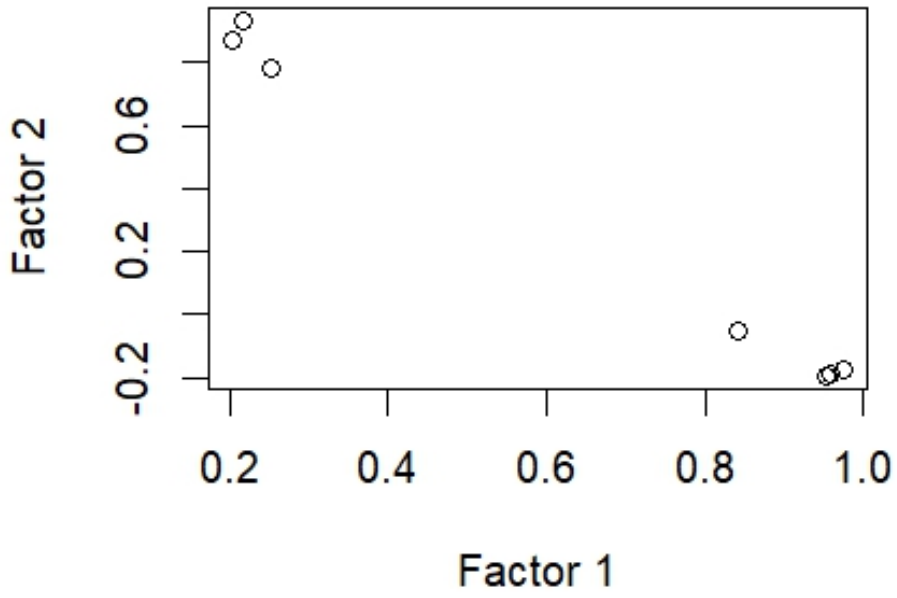
$$\hat{\mathbf{L}} = \begin{bmatrix} 0.9738 & -0.1744 \\ 0.9587 & -0.1885 \\ 0.8410 & -0.0513 \\ 0.9521 & -0.1920 \\ 0.2175 & 0.9282 \\ 0.2040 & 0.8676 \\ 0.2526 & 0.7828 \end{bmatrix}$$

Observe que **las cuatro primeras variables** (Sabor, Aroma, Color, Textura) **están fuertemente correlacionadas con** F_1, mientras que, **las 3 últimas** (Utilidad, Facilidad de enontrar, Empaque) **están mas correlacionadas con** F_2.

Análisis Factorial

Podemos ilustrar esto en un gráfico:

```
plot(L_m, xlab = "Factor 1", ylab = "Factor 2")
```



Análisis Factorial

Vamos a calcular la Matriz de Comunalidad:

$$\mathbf{H} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{\top}$$

```
# --- Matriz de Comunalidad
H<-L_m%*%t(L_m)
round(diag(H),4)
[1] 0.9787 0.9547 0.7099 0.9434 0.9089 0.7943 0.6766
```

Note que los dos factores logran explicar el 97.87 %, 95.47 %, 70.99 %, 94.34 %, 90.89 %, 79.43 % y 67.66 % de la variabilidad del sabor, aroma, color, textura, utilidad, localidad y empaque, respectivamente.

Análisis Factorial

La Matriz de Varianzas Específica:

$$\hat{\boldsymbol{\Psi}} = \text{Diag}\left(\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\top}\right)$$

```
# --- Matriz de Varianzas Específica
Psi<-diag(R-H)  # solo la diagonal de Psi
round(Psi,4)
[1] 0.0213 0.0453 0.2901 0.0566 0.0911 0.2057 0.3234
```

Observe que la variabilidad que no fue explicada por lo factores en cada una de las variables es **relativamente baja**. Las más altas se presentaron en las variables empaque (0.3234), color (0.2901) y localización (0.2057).

Análisis Factorial

También calculamos la **Matriz Residual**:

$$\text{MRES} = \mathbf{R} - \left(\hat{\mathbf{L}}\hat{\mathbf{L}}^{\top} + \hat{\boldsymbol{\Psi}}\right)$$

```
# --- Matriz Residual
Psi_completa<-diag( diag(R-H),ncol=7)
MRES<-R-(H+Psi_completa)
ECM<-sqrt ( sum(MRES*MRES)/42) # Error Cuadrático Medio
round(ECM,4)
[1] 0.0604
```

Análisis Factorial

Interpretación de los factores:

- En F_1 las mayores cargas son de las variables: **sabor, aroma , color y textura**. Así que este factor está representando o relacionando **la parte comestible del producto**. Por lo tanto, podemos pensar o ver a este factor como **un índice de la de opinión del consumidor en términos de los atributos comestibles del producto**. A mayor calificación de las 4 variables, mayor será la nota de éste factor (carga factorial).
- En F_2 las mayores cargas son las de **utilidad, facilidad para encontrar y empaque**. Este factor se puede ver como **la opinión del consumidor con respecto a la parte económica**. Entonces, cuanto más grande sea la nota del consumidor con respecto a éstas variables, más grande será la nota de éste factor.

Análisis Factorial

Podemos hacer el AF en R con:

```
# ----- principal
require(psych)
principal(R, nfactors=2,rotate="none", scores=TRUE)

Principal Components Analysis
Call: principal(r = R, nfactors = 2, rotate = "none", scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
   PC1   PC2   h2    u2 com
1 0.97 -0.17 0.98 0.021 1.1
2 0.96 -0.19 0.95 0.045 1.1
3 0.84 -0.05 0.71 0.290 1.0
4 0.95 -0.19 0.94 0.057 1.1
5 0.22  0.93 0.91 0.091 1.1
6 0.20  0.87 0.79 0.206 1.1
7 0.25  0.78 0.68 0.323 1.2

                       PC1  PC2
SS loadings           3.63 2.33
Proportion Var        0.52 0.33
Cumulative Var        0.52 0.85
Proportion Explained  0.61 0.39
Cumulative Proportion 0.61 1.00
```

2. Método de los Factores Principales Iterativo:

Se hace un **refinamiento** de $\mathbf{L}$ y $\mathbf{\Psi}$ obtenidas por el método de los factores principales.

Vimos que:

$$\mathbf{P} = \mathbf{L}\,\mathbf{L}^\top + \mathbf{\Psi}$$

de ahí

$$\mathbf{L}\,\mathbf{L}^\top = \mathbf{P} - \mathbf{\Psi}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} - \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

Análisis Factorial

$$\mathbf{L}\,\mathbf{L}^{\top} = \begin{bmatrix} 1-\psi_1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1-\psi_2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1-\psi_p \end{bmatrix} = \begin{bmatrix} h_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & h_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & h_p^2 \end{bmatrix}$$

donde $h_i^2 = 1 - \psi_i$, $i = 1, \ldots, p$ son las comunalidades.

Suponga que las ψ_i fueron previamente estimadas, es decir que **tenemos una estimación incial** de $\mathbf{\Psi}$, la cual denotamos por:

Análisis Factorial

$$\boldsymbol{\Psi}^* = \begin{bmatrix} \psi_1^* & 0 & \cdots & 0 \\ 0 & \psi_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p^* \end{bmatrix}$$

También que **estimamos P**:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Análisis Factorial

Luego,

$$\mathbf{L}^*\,\mathbf{L}^{*\top} \cong \begin{bmatrix} 1-\psi_1^* & r_{12} & \cdots & r_{1p} \\ r_{21} & 1-\psi_2^* & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1-\psi_p^* \end{bmatrix} = \underbrace{\begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^{*2} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & h_p^{*2} \end{bmatrix}}_{\mathbf{R}^*}$$

donde $h_i^{*2} = 1 - \psi_i^*$ son las respectivas **estimativas iniciales de las comunalidades** h_i^2, $i = 1, 2, \ldots, p$.

Análisis Factorial

Usando el método de las componentes principales tenemos que:

$$\mathbf{L}^* = \left[\sqrt{\hat{\lambda}_1^*}\hat{\mathbf{e}}_1^* \qquad \hat{\lambda}_m^*\hat{\mathbf{e}}_m^*\right]$$

donde $\hat{\lambda}_i^*$ son los **autovalores** de $\mathbf{R}^*$ y $\hat{\mathbf{e}}_m^*$ sus correspondientes **autovectores**, $i = 1, \ldots, m$.

A partir de $\mathbf{L}^*$ tenemos las nuevas estimativas de las comunalidades (h_i^*), las cuales son colocadas en la diagonal principal de $\mathbf{R}^* = \mathbf{L}^*\,\mathbf{L}^{*\top}$ y el proceso se repite hasta que las diferencias en las estimativas de las comunalidades de dos iteraciones sucesivas sean suficientemente pequeñas.

Análisis Factorial

¿Como obtenemos las estimativas iniciales de varianzas específicas?

Rta: Tenemos varias opciones. La más popular[2] es

$$\psi_i^* = \frac{1}{r^{ii}}$$

donde $r^{ii}(s^{ii})$ es el $i-$ésimo elemento de la diagonal de $\mathbf{R}^{-1}(o\,\mathbf{S}^{-1})$. Así

$$h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}}$$

que es igual al cuadrado del coeficiente de correlación múltiple entre X_i (variable respuesta) y las otras $p-1$ variables (variables explicativas).

[2] Cuando se trabaja con $\mathbf{R}$. Usamos s^{ii} cuando se trabaja con $\mathbf{S}$

Análisis Factorial

En resumen, el método se aplica de la siguiente forma:

- Paso 1 Calcule la matriz de correlación muestral $\mathbf{R}$.
- Paso 2 Obtenga las ψ_i^* (estimativas ψ_i), $i = 1, \ldots, p$.
- Paso 3 Determine $\mathbf{R}^*$ reemplazando $h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}}$ en la diagnonal principal de $\mathbf{R}$.
- Paso 4 Calcule $\hat{\lambda}_i^*$ y $\hat{\mathbf{e}}_i^*$, los autovalores y autovectores de $\mathbf{R}^*$.
- Paso 5 Determine m el número de factores.

Análisis Factorial

Paso 6 Calcule las comunalidades de cada variable con los m factores seleccionados

$$l_i^{*2} = l_{i1}^2 + \cdots + l_{im}^2$$

y las varianzas residuales (específicas)

$$1 - l_i^{*2}$$

Paso 7 Si las comunalidades y las especificidades son positivas, regrese al paso (3), utilice las nuevas estimaciones y repita nuevamente los pasos. Un ciclo formado por estos pasos concluye una iteración.

Observación: En algunas ocasiones, éste método no permite llegar a una estimación adecuada debido al problema de convergencia.

3. Método de Máxima Verosimilitud

Suponga que

- $\mathbf{X}_{p\times 1} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \longrightarrow \mathbf{Z} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{P})$
- $\mathbf{F} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I})$
- $\epsilon \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Psi})$

Si se observa muestra aleatoria $Z_1, \ldots, Z_n$ de $\mathbf{Z}$, la **función de verosimilitud** de $\mathbf{P}$ queda dada por

$$L(\mathbf{P} \mid \mathrm{z}) = \frac{1}{(2\pi)^{np/2} \mid \mathbf{P} \mid^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \mathrm{z}_j^{\top} \mathbf{P}^{-1} \mathrm{z}_j \right\}$$

Análisis Factorial

Reemplazando $\mathbf{P} = \mathbf{L}\,\mathbf{L}^\top + \mathbf{\Psi}$ tenemos

$$L(\mathbf{L}, \mathbf{\Psi} | \, z) = \frac{1}{(2\pi)^{np/2} | \, \mathbf{L}\,\mathbf{L}^\top + \mathbf{\Psi} |^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} z_j^\top [\mathbf{L}\,\mathbf{L}^\top + \mathbf{\Psi}]^{-1} z_j \right\}$$

Las estimaciones de $\mathbf{L}$ y $\mathbf{\Psi}$ son las matrices que maximizan $L(\mathbf{L}, \mathbf{\Psi} | \, z)$.

En general, esa maximización se hace numéricamente[3] o computacionalmente para un valor m fijado de antemano. El problema es que algunas veces el algoritmo puede no converger.

[3] Por ejemplo: Newton-Raphson

Análisis Factorial

Observaciones:

- ✓ ***Bajo la normalidad*** *de los datos, se recomienda el método de* ***máxima verosimilitud****, desde que* ***proporciona estimaciones más precisas****; en caso contrario (sin normalidad), se recomienda el método de factores principales.*
- ✓ *El método de máxima verosimilitud, exige el conocimiento de m.* ***Diferentes valores de*** *m* ***conducen a diferentes estimativas en las cargas factoriales****. Algunos autores recomiendan utilizar el método de componentes principales para determinar m y posteriormente el de máxima verosimilitud.*

Análisis Factorial

Example

Los datos del archivo `cereal.txt` corresponden a un estudio realizado por Roberts y Lattin (1991) en el cuál se investiga la percepción de los consumidores con respecto a sus marcas favoritas de cereal. Se le solicitó a cada entrevistado, evaluar (en una escala de 5 puntos) sus tres marcas favoritas, teniendo en cuenta 25 atributos.

Marcas: All Bran, Cerola Muesli, Just right, Kellogg's Corn Flakes, Komplete, NutriGrain, Purina Muesli, Rice Bubbles, Special K, Sustain, Vitabrit, Weetbrix.

Atributos: satisface, natural, fibra, dulce, facil, sal, gratificante, energia, divertido, nino, encharcado, economico, salud, familia, calorias, simple, crocante, regular, azucar, fruta, proceso, calidad, plazer, aburrido, nutritivo.

Análisis Factorial

```
 # ----- Cereal ----- #
Cereal <- read_table("D:/Desktop/Teaching/Dataset/Cereal.txt",
                     col_names = FALSE)

# --- Asignamos los nombres de los atributos
require(dplyr)
Cereal2<-select(Cereal,
                identidad=X1,numero=X2,satisface=X3,
                natural=X4, fibra=X5,dulce=X6,facil=X7,
                sal=X8, gratificante=X9,energia=X10,
                divertido=X11,nino=X12,encharcado=X13,
                economico=X14, salud=X15,familia=X16,
                calorias=X17, simple=X18,crocante=X19,
                regular=X20,azucar=X21, fruta=X22,
                proceso=X23,calidad=X24,plazer=X25,
                aburrido=X26,nutritivo=X27)
```

Análisis Factorial

```
X<-Cereal2[,3:27]    # Variables Analizar

R<-cor(X)            # Matriz de Correlacion
```

Primeramente verificamos si el análisis factorial es viable

```
# --- Prueba de esfericidad de Bartlett
require(psych)
cortest.bartlett(R)
$chisq
[1] 1153.931

$p.value
[1] 2.437858e-100
```

Análisis Factorial

observe que

$$p - \text{valor} = 2.437858e - 100 < 0.05 = \alpha$$

por lo tanto el **AF es viable**.

```
# --- KMO
KMO(R)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = R)
Overall MSA =  0.85
```

Observe que el criterio KMO también nos indica que el AF es viable.

Análisis Factorial

Para e AF, primero debemos **determinar el número de factores** m. Sabemos que bajo el método de los **factores principales**, la proporción de varianza explicada por F_j está dada por:

$$\text{PVE}_{F_j} = \frac{\lambda_i}{p}$$

```
auto<-eigen(R)
lambdai<-auto$values
e<-auto$vectors
PVE<-lambdai/sum(lambdai)
round(PVE,4)
 0.2602 0.1528 0.1001 0.0674 0.0434 0.0373 0.0341
 0.0315 0.0293 0.0278 0.0259 0.0219 0.0212 0.0196
 0.0167 0.0155 0.0145 0.0144 0.0122 0.0110 0.0105
 0.0097 0.0087 0.0079 0.0066
```

Análisis Factorial

Podemos hacer el AF en R con:

```
fac(R,
    nfactors = 3,      # No. de factores
    rotate = "none",   # sin rotación
    fm="ml")           # máxima verosimilitud

# fm    : metodo de estimacion de las cargas factoriales
#         "pa" - componentes principales
#         "ml" - maxima verosimilitud

# scores: metodo de estimacion de los scores factoriales
#         "regression" - via regresion (default)
#         "Bartlett"   - via componentes principales
```

Análisis Factorial

```
#                En la salida:
#PA o ML : cargas factoriales
#h2: Comunalidades
#U2: Especifidad
# SS loadings: autovalores de los factores (lambda)
# Proportion Var: prop. de var. exp. por cada fact.
```

Análisis Factorial

Una vez se hayan seleccionado los m **factores** y se tengas las **cargas factoriales** (peso de cada factor para cada variable) se pueden **estimar los scores** (valores numéricos) **de los factores** para cada elemento muestral y después utilizarlos para otros análisis (muestras).

Para cada elemento muestral:

$$\mathbf{F}_{jk} = w_{j1}Z_{1k} + w_{j2}Z_{2k} + \cdots w_{jp}Z_{pk}$$

$j = 1, 2, \ldots, n$. Esto parece/recuerda un **modelo de regresión lineal múltiple**. Así que podemos hacer las estimaciones de los pesos (w_{jp}) usando **mínimos cuadrados ponderados**:

$$\widehat{\mathbf{F}}_j = \left(\widehat{\mathbf{L}}^\top \widehat{\boldsymbol{\Psi}}^{-1} \widehat{\mathbf{L}}\right)^{-1} \left(\widehat{\mathbf{L}}^\top \widehat{\boldsymbol{\Psi}}^{-1}\right) \mathbf{z}_j = \mathbf{W}\,\mathbf{z}_j$$

Análisis Factorial

donde

- z_j es el vector de observaciones del individuo j $(j = 1, \ldots, n)$.
- $\mathbf{W}_{m \times p}$ es la matriz de ponderación que genera los coeficientes w_{ij}

Otro método para estimar los scores de $\mathbf{F}$ es el **método de regresión:**

$$\widehat{\mathbf{F}}_j = \widehat{\mathbf{L}}^\top \widehat{\mathbf{P}}^{-1} z_j$$

Análisis Factorial

Para los datos de `Cereal.txt`:

```
AF_CP<-fa(X,nfactors = 4,rotate = "varimax")
L<-AF_CP$loadings
print(AF_CP$loadings,cutoff = 0.6)
```

Podemos interpretar los **cuatro factores** como: **Saludable**, **Artificial (No saludable)**,**Familiar o Infantil**, **Interesante**.
Si queremos ordena las observaciones de acuerdo con el factor 1:
Cuando se calculen los scores de los factores, las observaciones que tengan los scores más altos en ese factor deben ser los cereales considerados como más saludables por los voluntarios.

Análisis Factorial

```
# ----- Ejemplo 9.5 de Johnson and Wichern (2013)
# ----- pp. 497-498
R<-matrix(c(1.000,0.632,0.511,0.115,0.155,
            0.632,1.000,0.574,0.322,0.213,
            0.511,0.574,1.000,0.183,0.146,
            0.115,0.322,0.183,1.000,0.683,
            0.155,0.213,0.146,0.683,1.000),
          ncol=5)

AF<-fac(R,nfactors = 2,rotate = "none",
        scores = "Bartlett", fm="ml")
```