# Análisis de Factores

**El análisis de factores o análisis factorial tiene como objetivo explicar un conjunto de variables observadas mediante un pequeño número de variables latentes, o no observadas (llamados factores), con la mínima pérdida de información**

**Ejemplo 1:**

- Supongamos que se toman 20 medidas físicas del cuerpo de una persona: longitud del tronco y de las extremidades, anchura de hombros, peso, etc.
- Las mediciones no son independientes entre sí, conocidas algunas de ellas se pueden prever las restantes con poco error.
- Esto se puede explicar por el hecho de que las dimensiones del cuerpo humano dependen de ciertos factores, que si son conocidos, se podrían prever las dimensiones con poco error.

**Ejemplo 2:**

- Se está interesado en estudiar el desarrollo humano en los países del mundo y se disponen de muchas variables económicas, sociales y demográficas.
- En general, todas las variables son dependientes entre si, que están relacionadas con el desarrollo.
- Podemos preguntarnos si el desarrollo de un país depende de un pequeño número de factores tales que, conocidos sus valores, podríamos prever el conjunto de las variables de cada país.

**Ejemplo 3:**

- Se mide la capacidad mental de un individuo para procesar información y resolver problemas mediante distintas pruebas.
- Existen algunos factores no directamente observables, que expliquen los resultados observados?
- El conjunto de estos factores será lo que llamamos *inteligencia*.
- Es importante conocer cuantas dimensiones distintas tiene este concepto y como caracterizarlas y medirlas.

**Ejemplo 4:**

- Se desea medir la capacidad de *abstracción*, *analítica* y *memoria* de los alumnos.
- Se observaron 10 calificaciones de cada alumno de un grupo de estudiantes universitarios.
- Entre estas notas, o al menos entre algunas de ellas, se observan correlaciones elevadas que, en cierta medida, provienen de aptitudes globales del alumno que no se observan directamente:
  *Áreas de evaluación*:

| | |
|---|---|
| Álgebra | Contabilidad financiera |
| Cálculo | Análisis de costos |
| Estadística | Comunicación comercial |
| Derecho mercantil | Actuariales |
| Derecho laboral | Econometría |

**Ejemplo 4:**

- Un análisis factorial podría permitir que la información relativa a estas variables se resumiese en tres únicos factores, sin pérdida excesiva de información
- Cada uno de estos tres factores se interpretaría como:

$F1$ – Factor de CAPACIDAD DE ABSTRACCIÓN

$F2$ – Factor de MEMORIA

$F3$ – Factor de CAPACIDAD ANALÍTICA

**El primer modelo de factores fue propuesto por Karl Pearson y Charles Spearman, en su interés por comprender las dimensiones de la inteligencia humana.**

# Análisis de Factores: origen

- Karl Pearson y Charles Spearman derivaron el análisis de factores bajo el siguiente argumento. Supongamos que las variables se pueden agrupar de acuerdo a sus correlaciones, es decir, supongamos que todas las variables dentro de un grupo particular están altamente correlacionadas entre ellas, pero tienen una correlacion pequeña con variables en un grupo distinto.
- Entonces es posible que cada grupo de variables represente una *única dimensión subyacente o factor*, que es responsable de las correlaciones observadas en cada grupo.
- Este tipo de estructura es lo que el análisis de factores busca identificar.

# Análisis de Factores: origen relacionado con el área de psicología

Primeros trabajos de referencia del Análisis de factores

- Holzinger, K. J. and Swineford, F. (1939). A study in Factor Analysis: The stability of a Bi-Factor solution, University of Chicago: Supplementary Educational Monorgraphs, Number 48.
- Spearman, C. (1904). General intelligence objectively determined and measured. American Journal of Psychology, 15, 201-293.
- Thurstone, L. L. (1935). Vectors of the mind. Chicago: University of Chicago Press.
- Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

# Objetivos de la aplicación del Análisis de factores

- **Seleccionar un subconjunto de variables** entre un gran número de opciones, con base en la selección de las más altas correlaciones entre las variables originales y los factores.
- **Generar un grupo de factores no correlacionados** como una forma de enfrentar la presencia de colinealidades en procedimientos de regresión múltiple.
- **Identificar grupos de individuos**
- **Identificar valores extremos o outliers**
- **Predecir valores faltantes**

# Análisis de Factores y Análisis de componentes principales (PCA)

- El análisis de factores se puede considerar una extensión de componentes principales.
- Ambas técnicas intentan reducir la dimensionalidad de los datos mediante un número pequeño de términos, de tal forma que éstos capturen de manera aproximada la variabilidad de los datos originales.
- Sin embargo, la manera en la que AF construye estos términos es mas elaborada. Al igual que PCA, Análisis de factores intenta explicar las correlaciones entre variable, pero considerando que existe una fuente de variación común a las variables y otra fuente de variación específica de cada variable.
- PCA es una herramienta descriptiva, mientras que AF presupone un modelo estadístico formal de generación de datos.

# Enfoques de aplicación del análisis de factores

1. **Análisis de factores exploratorio**
   - Busca descubrir la estructura *subyacente* de un gran número de variables mediante un conjunto de factores no observables
   - No se conoce a priori el número de factores. El investigador supone a priori que existen indicadores que estarán asociados a algún factor.
   - No existe una teoría a priori para identificar estas asociaciones sino que se interpretan las cargas de los factores para descubrir esas estructuras
2. **Análisis de factores confirmatorio**
   - Se asume que existen ciertos factores que explican la estructura de los datos de acuerdo a una teoría pre-establecida.
   - Las variables indicadoras se seleccionan con base en la teoría preestablecida
   - Se busca determinar o confirmar si los factores y sus cargas corresponden con lo que establece la teoría
   - El nombre y número de los factores puede definirse a priori.

# Areas de aplicación de análisis de factores

- Es una de las técnicas multivariadas más utilizadas y con más literatura disponible.
- Su aplicación es muy amplia, cubriendo áreas como mercadotecnia, psicología, educación, procesos de producción, salud, recursos humanos, desarrollos de nuevos productos, etc.

**Algunas referencias de la aplicación de análisis de factores:**

- Pett, Marjorie A., Nancy R. Lackey, and John J. Sullivan (2003). Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Thousand Oaks, CA: Sage Publications
- Widaman, K. F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters?"Multivariate Behavioral Research 28: 263-311. Cited with regard to preference for PFA over PCA in confirmatory factor analysis in SEM.

# Variabilidad en Mediciones

- Al igual que PCA, los modelos de análisis de factores se construyen basados en la variabilidad observada en las mediciones de variables.
- Sin embargo en AF se asume que la variabilidad de cada variable $X_i$ se descompone en dos términos: la varianza común y la varianza única (específica).
  - *La varianza común* es la parte de la variación de la variable que es compartida con las otras variables.
  - *La varianza única (específica)* es la parte de la variación de la variable que es propia de esa variable.
- Generalmente se refiere al AF como el análisis que busca un nuevo conjunto de variables, menor en número que las variables originales, que exprese lo que es común a esas variables e identificando también su variabilidad específica.

# Etapas del análisis de factores

1. **Recolección de datos y generación de matriz de covarianzas/correlaciones**. No es necesario disponer de los datos originales, la matriz de covarianzas/correlaciones es suficiente
2. **Obtención de factores**. Se obtienen las combinaciones lineales que forman los factores ortogonales, sin embargo los factores pueden ser no interpretables.
3. **Realizar una rotación de los factores para una interpretación adecuada**. Se busca una transformación que *cambie* los coeficientes de manera que facilite la interpretación de los factores. Existen varios métodos para encontrar la rotación más conveniente.
4. **Generación de escalas o puntajes de factores (factor scores)** para utilizarlos en análisis adicionales.

# Modelo de factores ortogonales

- Consideremos un vector aleatorio observable $\boldsymbol{X}$, con $p$ componentes, que tiene vector de medias $\mu$ y matriz de covarianzas $\boldsymbol{\Sigma}$.
- El modelo de factores establece que las variables observables de $\boldsymbol{X}$ son generadas mediante una combinación lineal de $m$ variables aleatorias no observables

$$F_1, F_2, \ldots, F_m (m < p)$$

llamadas *factores comunes* y $p$ fuentes de variación adicionales

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$$

llamados *errores* o también *factores específicos*

# Modelo de factores ortogonales

Algebraicamente el modelo de análisis de factores se establece de la siguiente forma:

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\
X_3 - \mu_3 &= l_{31}F_1 + l_{32}F_2 + \cdots + l_{3m}F_m + \varepsilon_3 \\
&\vdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

O bien en notación matricial

$$\boldsymbol{X} - \mu = \boldsymbol{L}\boldsymbol{F} + \varepsilon$$

# Modelo de factores ortogonales

- $\boldsymbol{L}$ representa la matriz de cargas de los factores y cada una de sus entradas, $l_{ij}$, representa la carga de la $i$-esima variable en el $j$-esimo factor
- El $i$-esimo factor específico $\varepsilon_i$ está asociado solo con la $i$-esima respuesta $X_i$
- Los $p$ elementos del vector de desviaciones $(\boldsymbol{X} - \mu)$ están expresados en términos de $p + m$ variables aleatorias

$$F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$$

las cuales son *no observables*

- Está característica distingue al modelo de análisis de factores del modelo de regresión lineal múltiple:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

donde $\boldsymbol{X}$ es un vector de variables observables.

# Modelo de factores ortogonales

Además se consideran los siguientes supuestos acerca de los vectores aleatorios $\boldsymbol{F}$ y $\varepsilon$:

$$E(\boldsymbol{F}) = \mathbf{0} \quad Cov(\boldsymbol{F}) = E(\boldsymbol{F}\boldsymbol{F}^{'}) = \boldsymbol{I}$$
$$E(\varepsilon) = \mathbf{0} \quad Cov(\varepsilon) = E(\varepsilon\varepsilon^{'}) = \boldsymbol{\Psi}$$

donde

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

$\boldsymbol{F}$ y $\varepsilon$ son independientes, es decir,

$$Cov(\varepsilon, \boldsymbol{F}) = E(\varepsilon\boldsymbol{F}^{'}) = E(\varepsilon)E(\boldsymbol{F}^{'}) = \mathbf{0}.$$

Así que la ecuación matricial del modelo, *junto* con estos supuestos constituyen el modelo de *factores ortogonales*

# Modelo de factores ortogonales

De los supuestos anteriores, podemos expresar la matriz de covarianzas de $\boldsymbol{X}$ como sigue:

$$\begin{aligned}\boldsymbol{\Sigma} &= Cov(\boldsymbol{X}) \\ &= E[(\boldsymbol{X}-\mu)(\boldsymbol{X}-\mu)^{'}] \\ &= E[(\boldsymbol{LF}+\varepsilon)(\boldsymbol{LF}+\varepsilon)^{'}] \\ &= E[\boldsymbol{LFF}^{'}\boldsymbol{L}^{'}+\boldsymbol{LF}\varepsilon^{'}+\varepsilon\boldsymbol{F}^{'}\boldsymbol{L}^{'}+\varepsilon\varepsilon^{'}] \\ &= \boldsymbol{LL}^{'}+\boldsymbol{\Psi}\end{aligned}$$

Con los mismos supuestos, se puede también concluir que:

$$\begin{aligned}Cov(\boldsymbol{X},\boldsymbol{F}) &= E[(\boldsymbol{X}-\mu)(\boldsymbol{F}-\boldsymbol{0})^{'}] = E[(\boldsymbol{X}-\mu)\boldsymbol{F}^{'}] \\ &= E[(\boldsymbol{LF}+\varepsilon)\boldsymbol{F}^{'}] = E[\boldsymbol{LFF}^{'}+\varepsilon\boldsymbol{F}^{'}] \\ &= E[\boldsymbol{LFF}^{'}] \\ &= \boldsymbol{L}\end{aligned}$$

# Modelo de factores ortogonales

De la relación

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{'} + \boldsymbol{\Psi}$$

Las varianzas y covarianzas entre las $X_i$ se pueden expresar por las cargas y las varianzas de los factores específicos (errores) de la siguiente forma:

$$Var(X_i) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i$$
$$Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{im}l_{km}$$

Se puede ver que cada covarianza entre una variable original y un factor común es la carga entre ellos

$$Cov(X_i, F_j) = l_{ij}$$

# Modelo de factores ortogonales

La varianza de la $i$-ésima variable se puede separar en dos partes:

1. La contribución de los factores comunes, expresada por la suma de cuadrados de las cargas en esa variable, que se denomina **comunalidad**.
2. La contribución del factor específico, llamada **varianza específica**.

$$\begin{aligned} Var(X_i) &= \sigma_{ii} \\ &= l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i \\ &= h_i^2 + \psi_i \end{aligned}$$

$$Comunalidad : h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$$

$$\text{Varianza específica} : \psi_i$$

# Modelo de factores ortogonales

- El modelo de factores asume que las $p + p(p-1)/2 = p(p+1)/2$ varianzas y covarianzas para $\boldsymbol{X}$ se pueden reproducir a partir de las $pm$ cargas $l_{ij}$ y de las $p$ varianzas específicas $\psi_i$.
- Cuando el número de factores es igual al número de variables, es decir, $m = p$, cualquier matriz de covarianzas $\boldsymbol{\Sigma}$ se puede reproducir exactamente como $\boldsymbol{LL}'$, de tal modo que $\boldsymbol{\Psi}$ sería una matriz de ceros, lo cual resulta poco práctico, debido a que la idea es reducir la dimensionalidad original de los datos
- Cuando $m = 1$, en muchas situaciones las matrices de covarianzas $\boldsymbol{\Sigma}$ no pueden ser factorizadas como $\boldsymbol{LL}' + \boldsymbol{\Psi}$, y en caso de que si se pueda obtener esta factorización, las soluciones no serán consistentes con la interpretación de los resultados.

# Rotación de los factores y cargas

- Cuando $m > 1$, siempre existirá alguna ambigüedad asociada con el modelo de factores. Si $\boldsymbol{T}$ es cualquier matriz ortogonal $m \times m$, es decir, $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$, entonces el modelo se puede escribir como

$$\begin{aligned} \boldsymbol{X} - \mu &= \boldsymbol{L}\boldsymbol{F} + \varepsilon \\ &= \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \varepsilon \\ &= \boldsymbol{L}^*\boldsymbol{F}^* + \varepsilon \end{aligned}$$

- Es decir, el modelo puede ser re-expresado a través de una rotación de las cargas y factores, estas rotaciones están dadas por:

$$\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{T} \quad \text{y} \quad \boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F}$$

$$E(\boldsymbol{F}^*) = \boldsymbol{T}' E(\boldsymbol{F}) = \boldsymbol{0}$$

$$Cov(\boldsymbol{F}^*) = \boldsymbol{T}' Cov(\boldsymbol{F})\, \boldsymbol{T} = \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}_{m \times m}$$

# Rotación de los factores y cargas

- Por tanto el nuevo factor $\boldsymbol{F}^*$ cumple los supuestos originales del modelo.
- Por tanto, la estructura de covarianza original $\boldsymbol{\Sigma}$ no se ve afectada por las rotaciones de las cargas y de los factores.
- Se busca una rotación adecuada que proporcione una fácil interpretación de los factores.
- Es imposible distinguir los pesos $\boldsymbol{L}$ de los $\boldsymbol{L}^*$ basado en las observaciones $\boldsymbol{X}$

$$\boldsymbol{F} \quad \text{y} \quad \boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F} \quad \text{tienen las mismas propiedades estadísticas}$$

- A pesar que las cargas $\boldsymbol{L}^*$ son en general diferentes de las cargas $\boldsymbol{L}$, ambas generan la misma matriz $\boldsymbol{\Sigma}$, debido a que

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{L}' + \boldsymbol{\Psi} = (\boldsymbol{L}^*)(\boldsymbol{L}^*)' + \boldsymbol{\Psi}$$

# Rotación de los factores y cargas

- Esta ambigüedad justifica la rotación de factores, debido a que las matrices ortogonales representan rotaciones del sistema de coordenadas para $\boldsymbol{X}$

**En resumen:**

- Las matrices $\boldsymbol{L}^*$ y $\boldsymbol{L}$ dan la misma representación de $\boldsymbol{\Sigma} = \boldsymbol{Cov}(\boldsymbol{X})$
- Las comunalidades dadas por los elementos de la diagonal de $\boldsymbol{LL}' = (\boldsymbol{L}^*)(\boldsymbol{L}^*)'$ no son afectadas por la elección de $\boldsymbol{T}$.
- Una vez obtenidas las cargas y los factores específicos es común obtener estimaciones para cada uno de los factores en cada uno de los $n$ casos, conocidos como *factor scores*

# Conveniencia del análisis de factores

- Dado un conjunto de $n$ observaciones vectoriales $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}$ en $p$ variables
- ¿Cuándo es apropiado utilizar un modelo con un número pequeño de factores para representar los datos?
- Si los elementos fuera de la diagonal de $\boldsymbol{S}$ son pequeños, o bien los elementos fuera de la diagonal de la matriz de correlaciones $\boldsymbol{R}$ son esencialmente cero, las variables no se relacionan entre sí y un análisis de factores no será de gran utilidad.
- Bartlett propuso una prueba para determinar si las variables no estan correlacionadas, denominada *prueba de esfericidad de Bartlett*
- El estadístico de prueba está dado por

$$-\left[(n-1)-\frac{(2p+5)}{6}\right] ln|\boldsymbol{R}| \sim \chi^2_{(p^2-p)/2}$$

# Prueba de Esfericidad de Bartlett

- El determinante de la matriz de correlaciones $\boldsymbol{R}$ es igual al producto de sus $p$ valores propios.
- Cuando $\boldsymbol{R}$ es próxima a la matriz identidad su determinante es próximo a 1 y por lo tanto su logaritmo natural es cercano a cero.
- Cuando $\boldsymbol{R}$ tiene valores propios próximos a cero, el determinante de $\boldsymbol{R}$ es próximo a cero y su logaritmo natural es un valor negativo con magnitud grande.
- Valores grandes del estadístico rechazan la hipótesis nula de que la matriz de correlaciones es igual a la matriz identidad.
- Se busca rechazar la hipótesis de esfericidad para proseguir con un análisis de componentes principales o análisis de factores.

# Métodos de estimación de los parámetros del modelo de factores

Existen varios métodos para estimar los parámetros del modelo, los métodos mas utilizados son:

1. Análisis de Componentes Principales(PCA, incluyendo el método del factor principal)
   - Método del factor principal: cuando se desea identificar variables latentes que contribuyan a la varianza común de las variables medidas, excluyendo la *varianza específica* (única).
2. El Método de máxima verosimilitud, asumiendo normalidad de los datos

Es aconsejable aplicar los dos métodos de estimación, las soluciones deberían ser consistentes unas con otras. Ambos métodos requieren cálculos iterativos que deben ser hechos por computadora.

# Estimación por componentes principales

- Proporciona una solución única que permite reconstruir las varianzas y covarianzas de las variables originales a partir de $p$ factores resultantes.
- Se obtienen tantos factores como variables originales.
- A través de diferentes criterios se elegirán los factores a considerar relevantes.
- Se utiliza generalmente cuando el objetivo es la reducción del número de componentes.
- Reproduce tanto la varianza común como la varianza específica de las variables en el estudio.

# Estimación por componentes principales

Se basa en la factorización de la matriz $\boldsymbol{\Sigma}$ mediante la descomposicion espectral. Sea $\boldsymbol{\Sigma}$ con $(\lambda_i, \boldsymbol{e_i})$ sus pares de valores y vectores propios tal que

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Entonces:

$$\Sigma = \lambda_1 \boldsymbol{e_1} \boldsymbol{e_1}' + \lambda_2 \boldsymbol{e_2} \boldsymbol{e_2}' + \cdots + \lambda_p \boldsymbol{e_p} \boldsymbol{e_p}'$$

$$= \left[ \sqrt{\lambda_1}\boldsymbol{e_1} \;\middle|\; \sqrt{\lambda_2}\boldsymbol{e_2} \;\middle|\; \cdots \;\middle|\; \sqrt{\lambda_p}\boldsymbol{e_p} \right] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e_1}' \\ \sqrt{\lambda_2}\boldsymbol{e_2}' \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{e_p}' \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}'$$

Por lo tanto, ajustamos la estructura de covarianza para el modelo de análisis de factores, teniendo tantos factores como variables $(m = p)$ y varianzas específicas $\psi_i = 0$ para toda $i$.

# Estimación por componentes principales

- La $j$-ésima columna de la matriz de cargas $\boldsymbol{L}$ está dada por $\sqrt{\lambda_j}\boldsymbol{e_j}$
- De esta forma se puede escribir

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \underset{(p\times p)}{\boldsymbol{L}}\underset{(p\times p)}{\boldsymbol{L}'} + \underset{(p\times p)}{\mathbf{0}} = \boldsymbol{L}\boldsymbol{L}'$$

- Además del factor escalar $\sqrt{\lambda_j}$, las cargas factoriales en el $j$-ésimo factor son los coeficientes de la $j$-ésima componente principal de la población.

# Estimación por componentes principales

- Lo que se busca es representar la estructura de la matriz varianzas y covarianzas $\boldsymbol{\Sigma}$ en términos de pocos factores comunes.
- Una aproximación se da cuando los $p-m$ valores propios son pequeños y entonces se elimina la contribución de

$$\lambda_{m+1}\boldsymbol{e_{m+1}}\boldsymbol{e_{m+1}}' + \lambda_{m+2}\boldsymbol{e_{m+2}}\boldsymbol{e_{m+2}}' + \cdots + \lambda_p \boldsymbol{e_p}\boldsymbol{e_p}'$$

de $\boldsymbol{\Sigma}$.

- Eliminando esa contribución, se obtiene la aproximación a $\boldsymbol{\Sigma}$ como

$$\boldsymbol{\Sigma} \doteq \left[ \sqrt{\lambda_1}\boldsymbol{e_1} \mid \sqrt{\lambda_2}\boldsymbol{e_2} \mid \cdots \mid \sqrt{\lambda_m}\boldsymbol{e_m} \right] \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e_1}' \\ \sqrt{\lambda_2}\boldsymbol{e_2}' \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e_m}' \end{bmatrix} = \underset{(p\times m)}{\boldsymbol{L}} \underset{(m\times p)}{\boldsymbol{L}'}$$

# Estimación por componentes principales

- Esta representación supone que los factores específicos $\varepsilon$ son de menor importancia y pueden ser eliminados de la factorización de $\mathbf{\Sigma}$.
- Si se incluyen los factores específicos en el modelo, sus varianzas pueden ser tomadas como los elementos diagonales de $\mathbf{\Sigma}$-$\boldsymbol{LL}'$ y la aproximación está dada por

$$\mathbf{\Sigma} \doteq \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e_1} & \sqrt{\lambda_2}\boldsymbol{e_2} & \cdots & \sqrt{\lambda_m}\boldsymbol{e_m} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e_1}' \\ \sqrt{\lambda_2}\boldsymbol{e_2}' \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e_m}' \end{bmatrix} + \begin{bmatrix} \widehat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \widehat{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\psi}_p \end{bmatrix}$$

donde $\widehat{\psi}_i = \sigma_{ii} - \sum_{j=1}^{m} \widehat{l}_{ij}^2$ para $i = 1, 2, .., p$

# Estimación por componentes principales

Para aplicar esta aproximación a una muestra de observaciones $\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n}$ es usual, primero, centrar las observaciones, restándoles la media muestral $\overline{\boldsymbol{X}}$

$$\boldsymbol{x_j} - \overline{\boldsymbol{X}} = \begin{bmatrix} X_{j1} - \overline{X}_1 \\ X_{j2} - \overline{X}_2 \\ \vdots \\ X_{jp} - \overline{X}_p \end{bmatrix}, \quad j = 1, ..., n$$

y en ocasiones se estandarizan

$$\boldsymbol{z_j} = \begin{bmatrix} \frac{X_{j\mathbf{1}} - \overline{X}_{\mathbf{1}}}{\sqrt{s_{\mathbf{11}}}} \\ \frac{X_{j\mathbf{2}} - \overline{X}_{\mathbf{2}}}{\sqrt{s_{\mathbf{22}}}} \\ \vdots \\ \frac{X_{jp} - \overline{X}_p}{\sqrt{s_{pp}}} \end{bmatrix}, \quad j = 1, ..., n$$

Cuya matriz de covarianzas es la matriz de correlaciones muestrales $\boldsymbol{R}$ de las observaciones $\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n}$

# Estimación por componentes principales

- La estandarización evita el problema de tener una variable con gran varianza, que afecta la determinación de las cargas de los factores.
- Entonces esta representación, cuando se aplican a la matriz $\boldsymbol{S}$ o a la matriz $\boldsymbol{R}$, se conoce como la solución del modelo de factores *por componentes principales*.
- El nombre proviene del hecho de que las cargas de los factores son los coeficientes *escalados* de unos cuantos componentes principales muestrales.

# Estimación por componentes principales

**Resultado:**

- El análisis de factores por componentes principales de la matriz de covarianzas $\boldsymbol{S}$ es especificado en términos de sus pares de valores y vectores propios $(\widehat{\lambda}_1, \widehat{\boldsymbol{e}}_{\boldsymbol{1}}), (\widehat{\lambda}_2, \widehat{\boldsymbol{e}}_{\boldsymbol{2}}), ..., \widehat{\lambda}_p, \widehat{\boldsymbol{e}}_{\boldsymbol{p}})$, donde

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$$

- Sea $m < p$, el número de factores comunes. Entonces la matriz de cargas estimada $\widetilde{\boldsymbol{L}}$ está dada por

$$\widetilde{\boldsymbol{L}} = \left[ \sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1 \;\middle|\; \sqrt{\widehat{\lambda}_2}\widehat{\boldsymbol{e}}_{\boldsymbol{2}} \;\middle|\; \cdots \;\middle|\; \sqrt{\widehat{\lambda}_m}\widehat{\boldsymbol{e}}_{\boldsymbol{m}} \right]$$

- Los vectores y valores propios obtenidos de $\boldsymbol{S}$, representan estimaciones de los verdaderos valores y vectores propios

# Estimación por componentes principales

**Resultado (Continuación)** Las varianzas específicas estimadas están dadas por los elementos de la diagonal de $\boldsymbol{S} - \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}'$, entonces:

$$\widetilde{\boldsymbol{\Psi}} = \begin{bmatrix} \widetilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \widetilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widetilde{\psi}_p \end{bmatrix} \quad \text{con} \quad \widetilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \widetilde{l}_{ij}^2$$

Y las comunalidades son estimadas por:

$$\widetilde{h}_i^2 = \widetilde{l}_{i1}^2 + \widetilde{l}_{i2}^2 + \cdots + \widetilde{l}_{im}^2$$

El análisis de factores por componente principales de la matriz de correlación muestral se obtiene empezando con $\boldsymbol{R}$ en lugar de $\boldsymbol{S}$

# Determinación del número de factores

- Para una solución por componentes principales, los estimadores de las cargas no cambian conforme aumenta el número de factores.
- Es decir, si se consideran por ejemplo 3 factores principales y se obtienen las cargas correspondientes, entonces si aumentamos el numero de factores a 4, las cargas de los primeros 3 factores no cambian.

**Cómo determinar el número de factores a utilizar?**
Si no se tienen consideraciones a priori, la elección de $m$ se puede basar en los valores propios estimados, de la misma manera como se hace con componentes principales

# Criterios para determinar el número de factores

- Idealmente, las contribuciones de los primeros factores a las varianzas muestrales de las variables deben ser grandes
- De manera análoga a componentes principales, la contribución a la varianza muestral total, $s_{11} + s_{22} + \cdots + s_{pp} = tr(\boldsymbol{S})$, del primer factor común es entonces:

$$\tilde{l}_{11}^2 + \tilde{l}_{21}^2 + \cdots + \tilde{l}_{p1}^2 = (\sqrt{\hat{\lambda}_1}\,\hat{\boldsymbol{e}}_1)'(\sqrt{\hat{\lambda}_1}\,\hat{\boldsymbol{e}}_1) = \hat{\lambda}_1$$

ya que los valores propios $\hat{\boldsymbol{e}}_1$ tienen longitud unitaria. En general

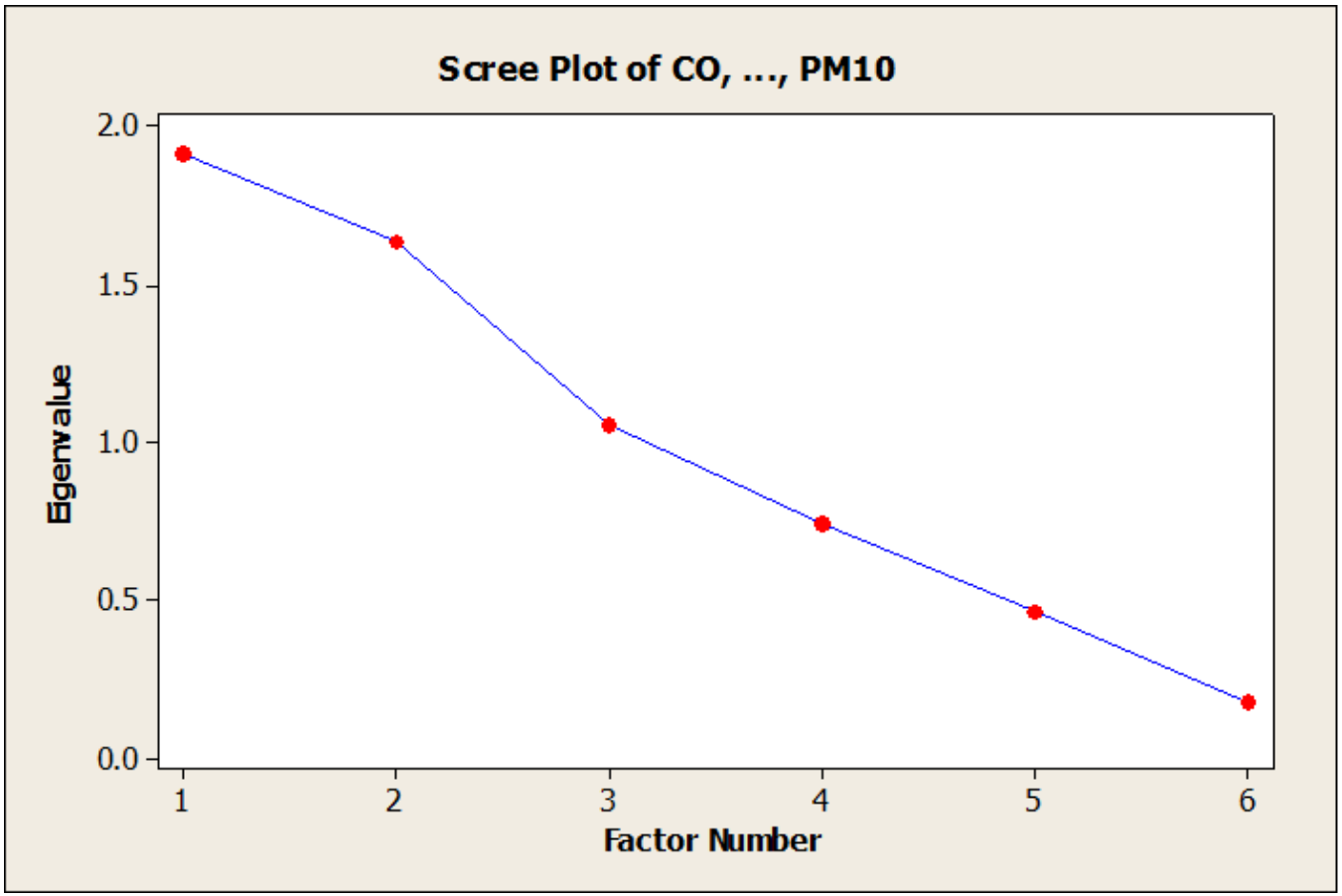
$$\left(\begin{array}{c}\text{Proporción de la}\\ \text{varianza muestral}\\ \text{total explicada}\\ \text{por el j-esimo factor}\end{array}\right) = \begin{cases} \dfrac{\hat{\lambda}_j}{s_{11}+s_{22}+\cdots+s_{pp}} & \text{para un análisis de factores de } \boldsymbol{S} \\ \dfrac{\hat{\lambda}_j}{p} & \text{para un análisis de factores de } \boldsymbol{R} \end{cases}$$

- Por lo general se elige el número de factores que explican de manera acumulada más del 70% de la varianza total

# Criterios para determinar el número de factores

- **Criterio de Kaiser**. Es uno de los criterios más utilizados por los paquetes estadísticos para determinar el número adecuado de factores. Se aplica cuando se trabaja con la matriz de correlaciones muestrales $\boldsymbol{R}$.
  - Consiste en elegir el número de factores como el número de valores propios de $\boldsymbol{R}$ mayores a 1.0
  - Esta regla busca que cualquier factor retenido contenga al menos la varianza de una de las variables utilizadas en el análisis.
  - Cuando el número de variables es pequeño, el número de valores propios $> 1$ será pequeño, por tanto esta regla puede detectar menos factores que los que existen realmente.
  - Cuando el número de variables es grande, esta regla puede detectar más factores que los que realmente existen.

# Criterios para determinar el número de factores



# Otros criterios para determinar el número de factores

**Parallel analysis (PA)**, conocido también como *Humphrey-Ilgen parallel analysis*.

- Algunos estudios lo refieren como el mejor método (Velicer, Eaton, and Fava, 2000: 67; Lance, Butts, and Michels, 2006).
- Mediante este método realiza un análisis de factores sobre una matriz de datos generada aleatoriamente y cuyas variables no están correlacionadas. Se considera el mismo número de casos y el mismo número de variables que tienen los datos reales.
- Se comparan gráficamente las dos líneas del scree plot, una para el análisis de los datos reales y otra para el análisis de los datos aleatorios.
- El número de factores elegido es aquel donde las dos líneas *se intersectan*.

# Otros criterios para determinar el número de factores

**Criterio de comparación de matrices de correlación**. Se obtiene la diferencia entre la matriz de correlaciones real y la matriz reproducida mediante $m$ factores, si el número de factores $m$ es adecuado, la diferencia debe ser cercana a cero.

- **Correlaciones reproducidas**. Es la matriz de correlaciones de las variables originales que resultaría suponiendo que es correcto el número de los factores retenidos.
- **Matriz de correlaciones residuales**. Es la matriz de las diferencias entre la matriz de correlaciones reproducida y la matriz de correlaciones reales. El numero adecuado de factores tendrá diferencias cercanas a cero.

El mejor enfoque es retener pocos en lugar de muchos factores, suponiendo que proporcionan una interpretación satisfactoria de los datos y dan un ajuste satisfactorio a ***S*** o ***R***.