laSalle
UNIVERSIDAD RAMON LLULL

MD004

Análisis de Factores

Máster Universitario en Data Science

Ricard Sierra Calls
Xavier Vilasís
Míriam Calvo

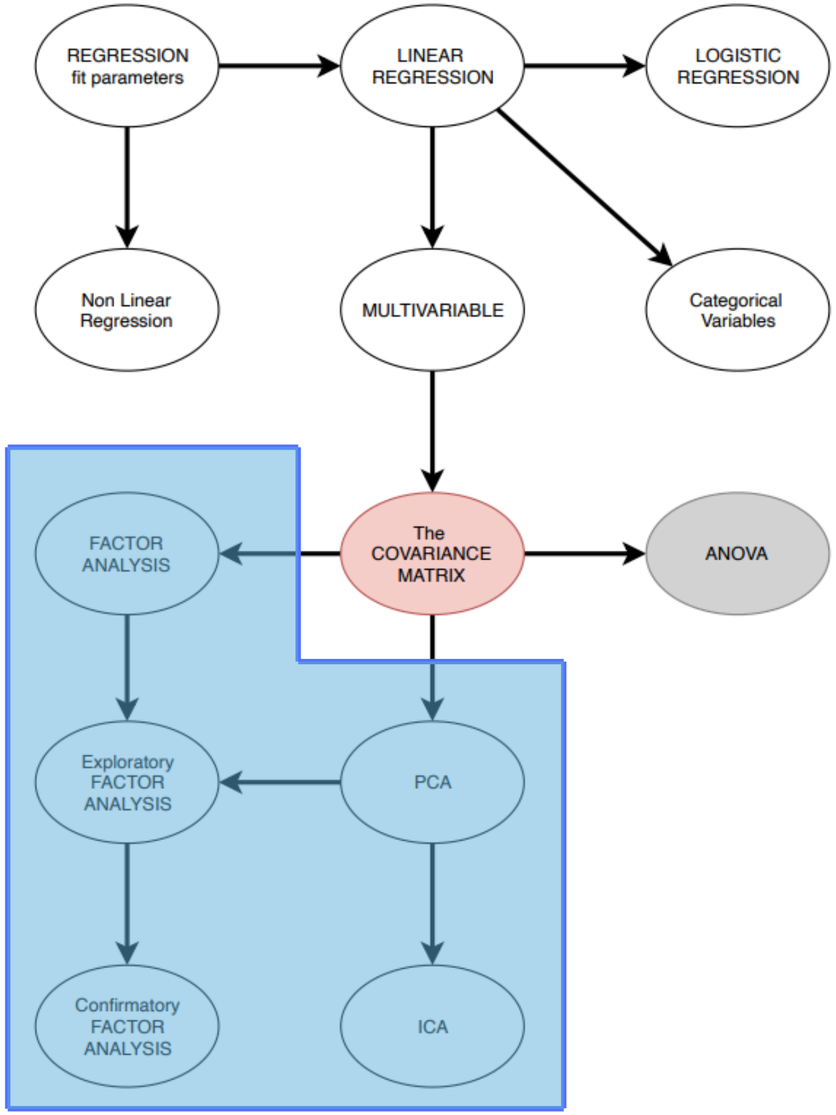
laSalle
UNIVERSIDAD RAMON LLULL

Índice

El **análisis de factores** es un método estadístico utilizado para describir la variabilidad entre variables observadas y correlacionadas en términos de un número potencialmente inferior de variables no observadas y no correlacionadas denominadas factores.

El análisis de factores busca esas variaciones conjuntas en respuesta a variables latentes no observadas. Las variables observadas se modelizan como combinaciones lineales de los factores potenciales, más los términos de ruido. La información obtenida sobre las interdependencias entre las variables observadas puede utilizarse posteriormente para reducir el conjunto de variables de un conjunto de datos.

Desde el punto de vista computacional, esta técnica equivale a una aproximación de bajo rango de la matriz de variables observadas.

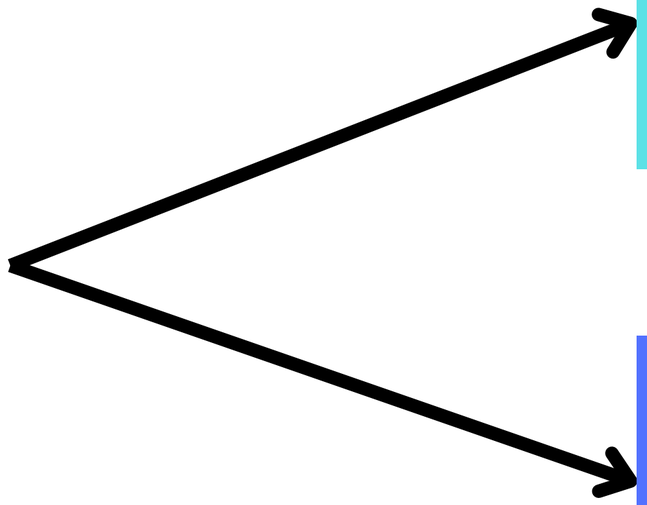


laSalle
UNIVERSIDAD RAMON LLULL

Tenemos un grupo de estudiantes y queremos evaluar su desempeño en diferentes aspectos académicos. Realizas un cuestionario con las siguientes variables observadas:

Variables que observamos

- Nota en Matemáticas.
- Nota en Física.
- Nota en Química.
- Nota en Literatura.
- Nota en Historia.



Factor 1: Habilidad lógica-matemática.

- Influirá fuertemente en matemáticas, física y química.

Factor 2: Habilidad verbal.

- Influirá más en literatura e historia.

laSalle
UNIVERSIDAD RAMON LLULL

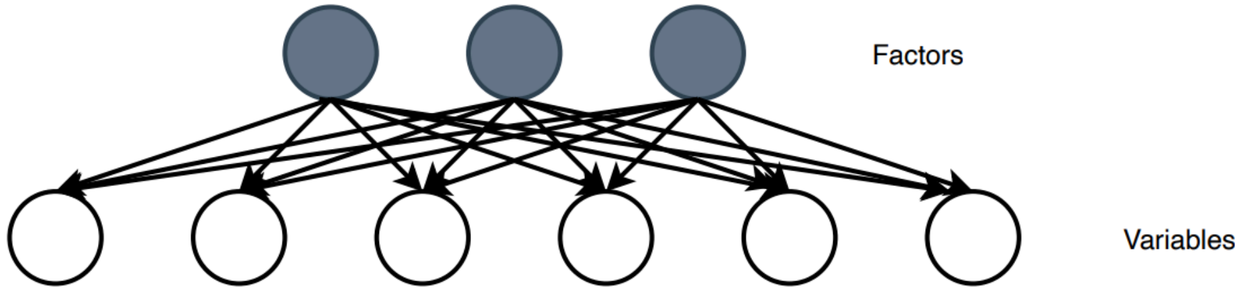
Tipos de análisis de factores

- **El análisis factorial exploratorio (AFE)** es un método estadístico utilizado para descubrir la estructura subyacente de un conjunto relativamente amplio de variables.
    - Ejemplo: Descubrir dimensiones en un conjunto de variables de cuestionarios

- **El análisis factorial confirmatorio (AFC)** es una forma especial de análisis factorial que se utiliza sobre todo en la investigación social. Se utiliza para comprobar si las medidas de un constructo son coherentes con la idea que tiene el investigador de la naturaleza de ese constructo (o factor). Como tal, el objetivo del análisis factorial confirmatorio es comprobar si los datos se ajustan a un modelo de medición hipotético. Este modelo hipotético se basa en la teoría y/o en investigaciones analíticas previas.
    - Ejemplo: Confirmar si las variables asociadas a la "satisfacción del cliente" realmente se agrupan en factores como "servicio", "calidad", etc.

- **Modelo Base**: $x = \Lambda f + \epsilon$
  - x: Vector de variables observadas.
  - Λ: Matriz de cargas factoriales (peso de cada factor en cada variable).
  - f : Vector de factores latentes (variables no observadas).
  - ϵ : Vector de ruido o error (variación no explicada por los factores).

- Las variables observadas (*x*) son una combinación de factores latentes (*f*) más algo de ruido (ϵ).
- **Analogía**: Imaginad que estais escuchando un coro de personas (*x*). Las voces individuales (*f*) son los factores, y el ruido (ϵ) es todo lo que no puedes identificar claramente.



- **Matriz de covarianza** de *x:* $\Sigma_x = \Lambda\Lambda^T + \Psi$
  - $\Lambda\Lambda^T$: Parte de la covarianza explicada por los factores.
  - Ψ: Varianza residual (ruido), asumida como una matriz diagonal (cada variable tiene un ruido independiente).

- Las covarianzas observadas entre las variables (*x*) se explican en gran parte por la influencia conjunta de los factores (*f*).
- El análisis de factores descompone la matriz de covarianzas para identificar cuáles factores son más relevantes.

- **Ejemplo**: Imaginad que tenéis datos de calificaciones de estudiantes (matemáticas, física, química) y observas que están correlacionadas. Los factores (habilidades subyacentes como "lógica matemática") explican por qué estas materias tienen alta covarianza.

laSalle
UNIVERSIDAD RAMON LLULL

Componentes del Modelo

- **Cargas factoriales (Λ)**:
  - Representan la relación entre factores y variables observadas.
  - Ejemplo: Si "habilidad matemática" es un factor, tendrá una carga alta en matemáticas y física, pero baja en literatura.

- **Comunalidad (h^2)**:
  - Porción de la varianza de una variable explicada por los factores:
    - **h^2** = **ΛΛ^2** (suma de las cargas al cuadrado para una variable).
  - Ejemplo: Si la habilidad matemática explica el 80% de la varianza en matemáticas, su comunalidad es 0.8.

- **Ruido (ϵ)**:
  - Varianza no explicada por los factores, que corresponde a los elementos diagonales de **Ψ**.

laSalle
UNIVERSIDAD RAMON LLULL

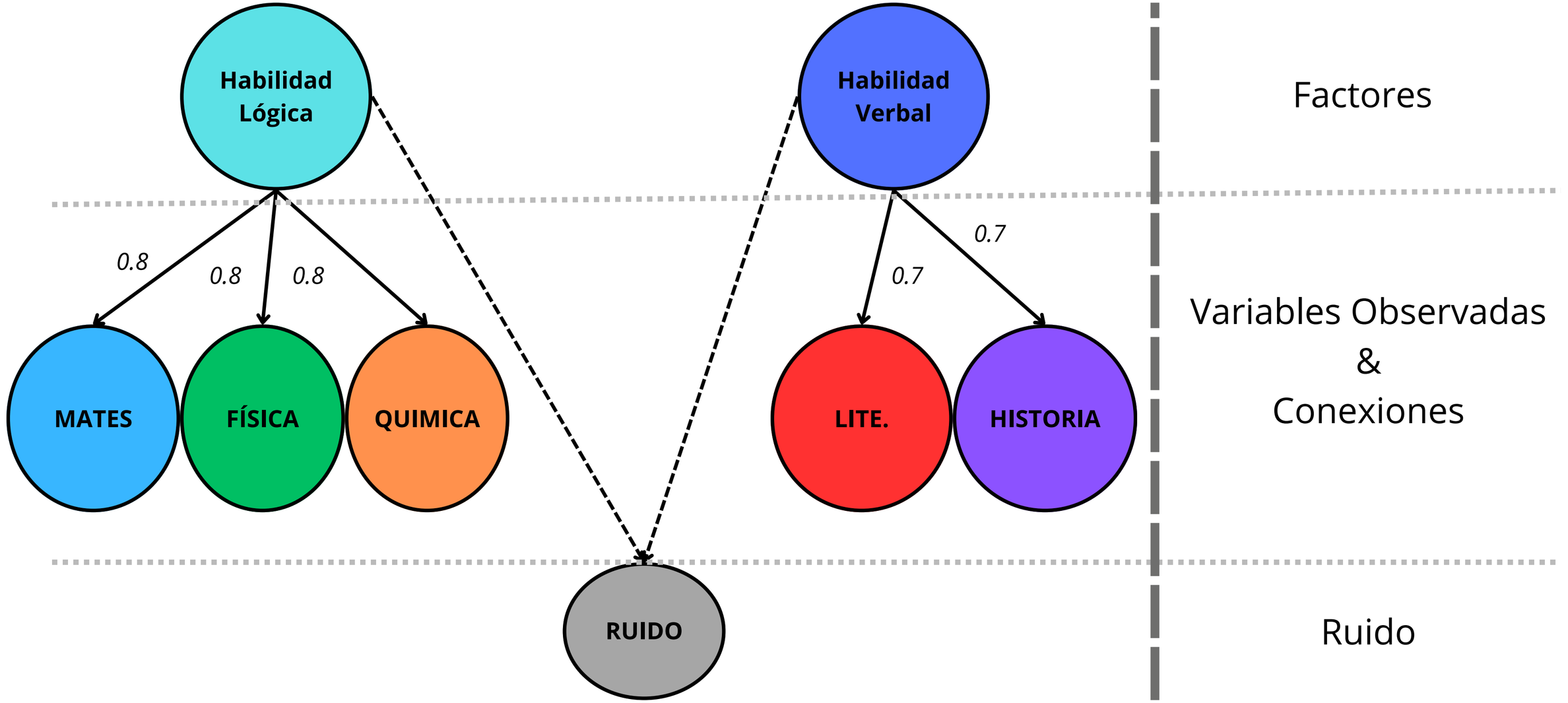
Suposiciones Clave

- **Independencia entre factores**:
  - Los factores no están correlacionados entre sí en el modelo estándar (aunque esto puede cambiar con rotaciones oblicuas).

- **Ruido independiente**:
  - Cada componente de ruido es independiente de los factores y de otros ruidos.

- **Factores estandarizados**:
  - Los factores tienen una varianza de 1.

laSalle
UNIVERSIDAD RAMON LLULL



laSalle
UNIVERSIDAD RAMON LLULL

Análisis de componentes principales (PCA)

Concepto clave

- Busca combinaciones lineales de las variables que maximizan la varianza.
- Las componentes principales están ordenadas por la cantidad de varianza explicada.

Fundamentos

- Basado en la matriz de covarianzas: Identifica las direcciones principales (autovectores) y las varianzas explicadas (autovalores).
- Gráfica del codo: Presenta la varianza acumulada y cómo seleccionar el número óptimo de componentes.

PCA identifica patrones generales, pero no asume un modelo estadístico subyacente, como el análisis de factores.

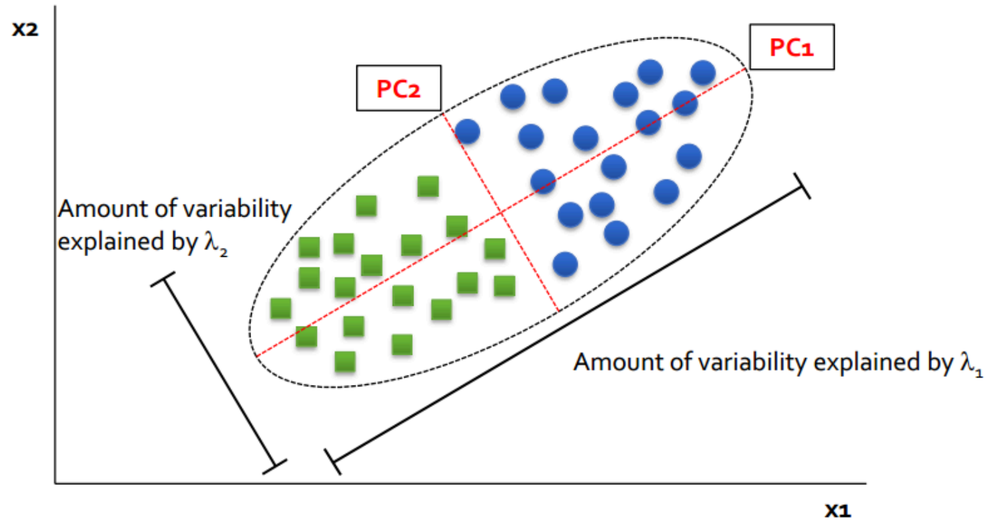
laSalle
UNIVERSIDAD RAMON LLULL

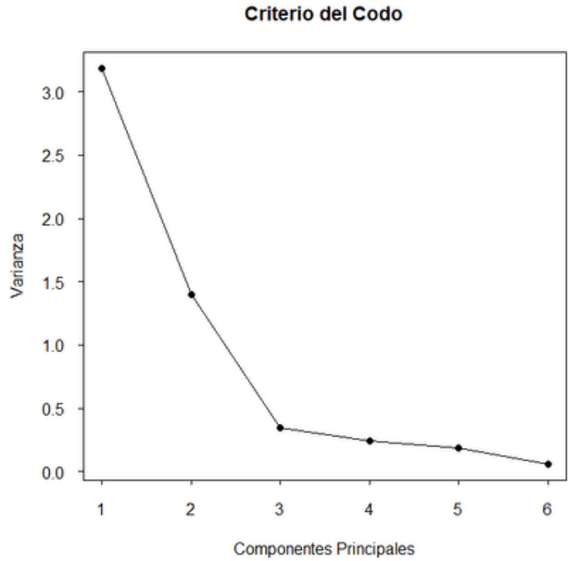
Análisis de componentes principales (PCA)

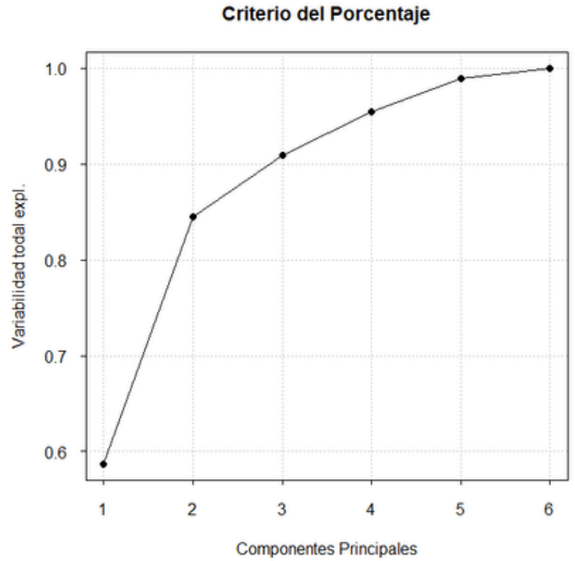
Aspecto	PCA	Análisis de Factores
Enfoque	Maximizar varianza	Modelar correlaciones
Modelo subyacente	No	Sí $(x = \Lambda f + \epsilon)$
Dimensionalidad	Componentes ortogonales	Factores latentes

laSalle
UNIVERSIDAD RAMON LLULL

Análisis de componentes principales (PCA)







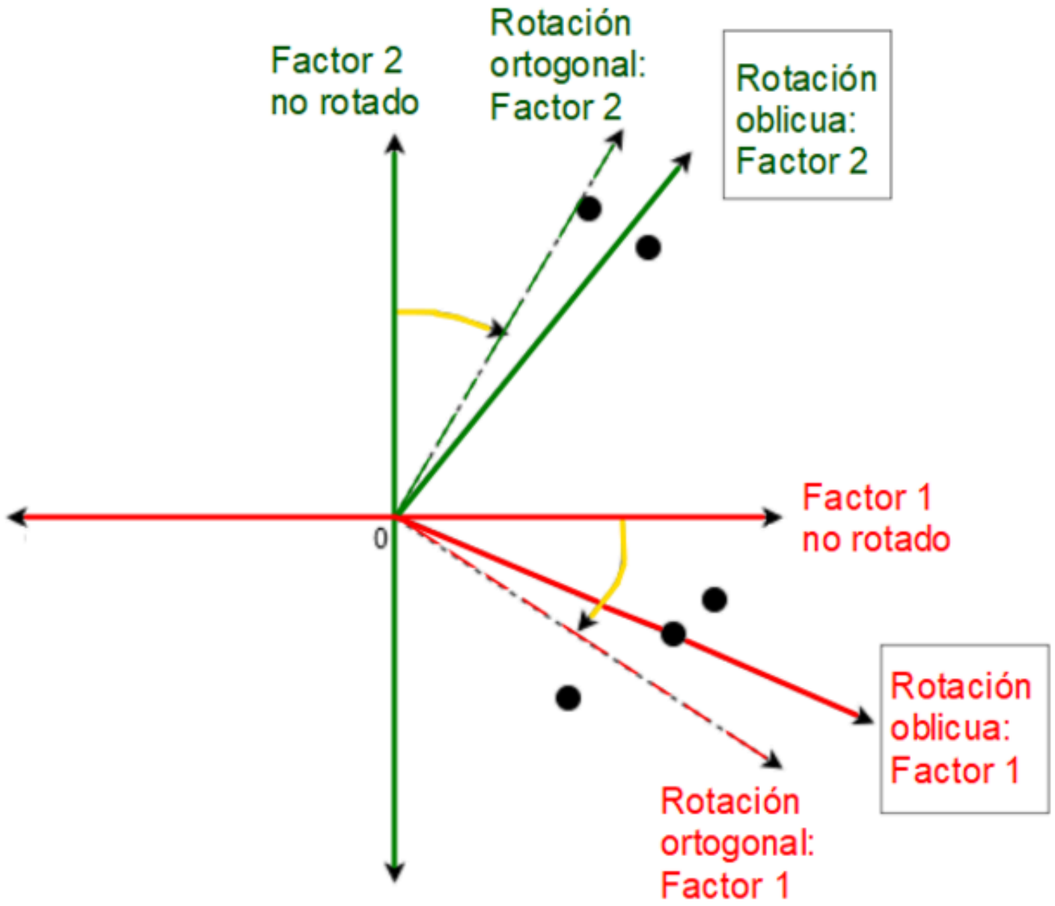
laSalle
UNIVERSIDAD RAMON LLULL

Tipos de Rotación y Su Impacto

- **Por qué rotar:**
  - Sin rotación, los factores pueden ser difíciles de interpretar porque las cargas factoriales están distribuidas en muchas variables.

- **Tipos de rotación:**
  - Varimax (ortogonal): Los factores permanecen no correlacionados. Útil para análisis exploratorios simples.
  - Promax (oblicua): Los factores pueden estar correlacionados. Mejor para datos más complejos.

Tipos de Rotación y Su Impacto



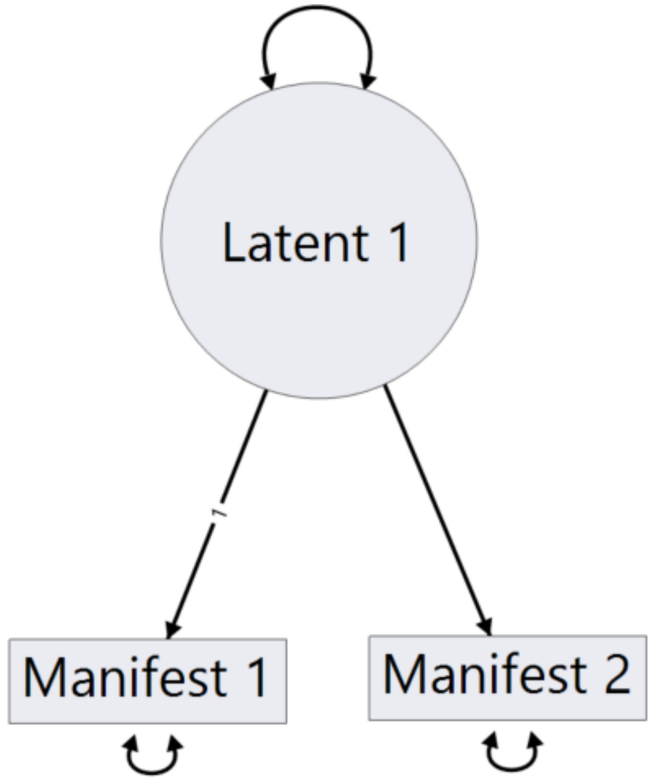
- **Rotación Ortogonal**: Los ejes de rotación forman un ángulo de 90 grados
- **Rotación Oblicua**: los ejes de rotación **no** forman un ángulo de 90 grados

Confirmación del Modelo

- **Confirmatory Factor Analysis (CFA):**
  - Se utiliza para validar un modelo predefinido basado en teoría.
    - Indicadores clave: GFI, CFI y RMSEA.

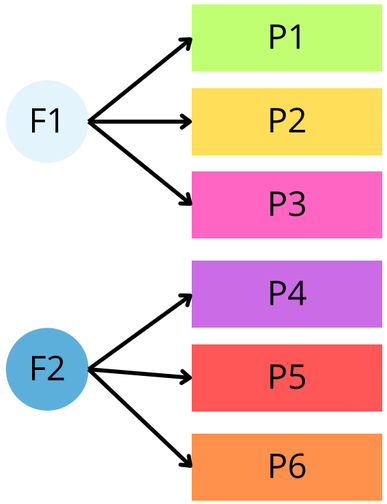
- **Relación con el análisis exploratorio:**
  - Mientras que el análisis exploratorio descubre patrones, el CFA confirma si esos patrones son consistentes con un modelo teórico.

- **Ejemplo práctico:**
  - Validar si un cuestionario de "satisfacción del cliente" realmente mide los factores "calidad del producto" y "atención al cliente".



laSalle
UNIVERSIDAD RAMON LLULL

Ejemplo CFA: Satisfacción del Cliente



Pregunta	Carga Factorial	Factor Asociado
P1	0.8	Calidad del Producto
P2	0.7	Calidad del Producto
P3	0.9	Calidad del Producto
P4	0.6	Atención al Cliente
P5	0.8	Atención al Cliente
P6	0.7	Atención al Cliente

- **GFI: 0.92** (bueno, >0.9).
- **RMSEA: 0.05** (bueno, <0.06).
- **Varianza explicada**:
  - (F1) Calidad del Producto: 75%.
  - (F2) Atención al Cliente: 70%.

- **Validación**: Las preguntas se agrupan bien en los factores correctos.
- **Punto fuerte**: P3 y P5 son las preguntas más representativas.
- **Mejora**: P4 tiene una carga baja (0.6); puede necesitar ajuste.

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 1: Examinar la Matriz de Correlaciones o Covarianzas

Paso 2: Selección del Número de Factores

Paso 3: Extracción Inicial de Factores

Paso 4: Rotación de Factores

Paso 5: Interpretación de Factores

Paso 6: Evaluación del Modelo

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 1: Examinar la Matriz de Correlaciones o Covarianzas

- **¿Por qué es importante?**
  - El análisis de factores se basa en la correlación entre variables observadas. Si las variables no están correlacionadas, el análisis no tiene sentido, ya que no habría patrones comunes que explicar.

- **Qué hacer:**
  - Calcula la matriz de correlaciones (para variables estandarizadas) o la matriz de covarianzas (para variables no estandarizadas).
  - Busca relaciones fuertes entre las variables:
    - Ejemplo: Si Matemáticas, Física y Química tienen correlaciones >0.7, es probable que compartan un factor común (e.g., Habilidad Lógica).

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 2: Selección del Número de Factores

- **¿Por qué es importante?**
  - No todas las correlaciones pueden explicarse por un solo factor. Este paso determina cuántos factores se necesitan para describir adecuadamente los datos.

- **Métodos comunes:**
  - Criterio de Kaiser:
    - Selecciona factores con autovalores mayores a 1 (cada factor explica más varianza que una variable individual).
  - Gráfica del codo (Scree Plot):
    - Grafica los autovalores en función del número de factores. Busca el "punto de inflexión" donde la pendiente se aplana (indicando factores que aportan poca varianza adicional).

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 3: Extracción Inicial de Factores

- **¿Por qué es importante?**
  - Este paso busca identificar los factores subyacentes que explican las correlaciones observadas entre las variables.

- **Métodos de extracción:**
  - Método de Componentes Principales (PCA):
    - Se utiliza como paso inicial para simplificar los datos. Las componentes principales identificadas se interpretan como factores iniciales.
  - Máxima verosimilitud:
    - Método estadístico que ajusta los factores de manera que maximicen la probabilidad de observar los datos dados los factores latentes.

- **Modelo inicial:**
  - Representa las variables observadas como combinaciones lineales de los factores más el ruido:

$$x = \Lambda f + \epsilon$$

  - La matriz de cargas factoriales (Λ) describe la influencia de cada factor en cada variable.

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 4: Rotación de Factores

- **¿Por qué es importante?**
  - Sin rotación, las cargas factoriales pueden ser difíciles de interpretar porque están distribuidas de forma uniforme en muchas variables.

- **Tipos de rotación:**
  - Ortogonal (Varimax):
    - Simplifica la interpretación al hacer que cada variable se relacione fuertemente con uno o pocos factores.
  - Oblicua (Promax):
    - Permite que los factores estén correlacionados, útil cuando los factores no son completamente independientes.

- **Ejemplo práctico:**
  - Antes de la rotación, Matemáticas, Física y Química podrían tener cargas similares en varios factores.
  - Después de la rotación, estas variables están claramente asociadas con un solo factor (Habilidad Lógica).

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 5: Interpretación de Factores

- **¿Cómo interpretamos los factores?**
  - Cada factor se interpreta en función de las variables que tienen altas cargas factoriales en ese factor.
  - Ejemplo:
    - Si el Factor 1 tiene altas cargas en Matemáticas, Física y Química, se interpreta como "Habilidad Lógica".

- **Comunalidad (h^2):**
  - Calcula la proporción de la varianza de cada variable explicada por los factores.
  - Ejemplo:
    - h^2=0.8: El 80% de la varianza en Matemáticas es explicada por los factores.

- **Ruido (ϵ):**
  - Evalúa la varianza no explicada para entender las limitaciones del modelo.

laSalle
UNIVERSIDAD RAMON LLULL

Proceso General del Análisis de Factores

Paso 6: Evaluación del Modelo

- **Criterios de un buen modelo:**
  - Los factores explican una proporción significativa de la varianza total (>60% es un estándar).
  - Las cargas factoriales son consistentes con la teoría o la intuición sobre los datos.

- **Revisión iterativa:**
  - Ajusta el número de factores o el método de extracción si el modelo no es satisfactorio.

laSalle
UNIVERSIDAD RAMON LLULL

¿Por qué es útil el análisis de factores?

- **Reducción de dimensionalidad:**
  - Simplifica datos complejos en variables más manejables.
- **Descubrimiento de patrones ocultos:**
  - Identifica relaciones subyacentes entre variables.
- **Validación de teorías:**
  - Confirma modelos conceptuales con datos observados.

laSalle
UNIVERSIDAD RAMON LLULL

Área de Aplicación	Ejemplo Específico	Variables Observadas	Factores Latentes	Objetivo
Psicología y Ciencias Sociales	Escala de personalidad Big Five (OCEAN)	Preguntas sobre comportamientos y actitudes	Apertura, Conciencia, Extroversión, Amabilidad, Neuroticismo	Diseñar y validar cuestionarios que midan rasgos de personalidad.
Marketing y Comportamiento del Consumidor	Encuesta de satisfacción del cliente	Opiniones sobre calidad, precio, atención al cliente	Calidad del producto, Experiencia de compra, Relación calidad-precio	Identificar aspectos clave que influyen en la satisfacción del cliente.
Ciencia de Datos y Machine Learning	Reducción de dimensionalidad en datos financieros	Ingresos, gastos en ocio, gastos en vivienda, ahorros	Nivel económico, Patrones de gasto	Simplificar datos para mejorar modelos predictivos o clustering.
Economía y Finanzas	Indicadores económicos compuestos	PIB, tasa de desempleo, inflación	Desarrollo económico, Estabilidad financiera	Construir índices que resuman múltiples indicadores en un solo valor.
Educación	Evaluación de habilidades académicas	Notas en Matemáticas, Física, Química, Literatura	Habilidad lógica-matemática, Habilidad verbal	Agrupar materias para evaluar habilidades subyacentes de los estudiantes.
Salud Pública y Medicina	Estudios epidemiológicos	Hábitos alimenticios, actividad física, sueño	Estilo de vida saludable, Riesgo de enfermedad	Identificar factores de riesgo y patrones de comportamiento.

laSalle
UNIVERSIDAD RAMON LLULL

Textura de alimentos

- Descripción: Medidas de la textura de un alimento pastelería.
- Fuente de datos:
  - 1.Aceite: porcentaje de aceite en la masa.
  - 2.Densidad: densidad del producto (cuanto mayor sea el número, más denso será el producto)
  - 3.Crujiente: medida de la textura crujiente, en una escala de 7 a 15, siendo 15 más crujiente.
  - 4.Fractura: el ángulo, en grados, con el que se puede doblar lentamente la masa antes de que se rompa.
  - 5.Dureza: se utiliza una punta afilada para medir la cantidad de fuerza necesaria antes de que se produzca la rotura.
- Datos simulados, pero con las características de un problema industrial.
- Forma de los datos: 50 filas y 5 columnas
- Restricciones de uso: Ninguna
- Persona de contacto: Kevin Dunn
- Datos de contacto: kgdunn@gmail.com
- Añadido aquí el 09 enero 2011 9:46
- Última actualización: 11 de noviembre de 2018 16:33

```
'data.frame': 50 obs. of 5 variables:
$ Oil : num 16.5 17.7 16.2 16.7 16.3 19.1 18.4 17.5 15.7 16.4 ...
$ Density : int 2955 2660 2870 2920 2975 2790 2750 2770 2955 2945 ...
$ Crispy : int 10 14 12 10 11 13 13 10 11 11 ...
$ Fracture: int 23 9 17 31 26 16 17 26 23 24 ...
$ Hardness: int 97 139 143 95 143 189 114 63 123 132 ...
```

```
Call:
factanal(x = food, factors = 1)

Uniquenesses:
     Oil  Density   Crispy Fracture Hardness
   0.616    0.524    0.045    0.255    0.850

Loadings:
          Factor1
Oil        0.620
Density   -0.690
Crispy     0.977
Fracture  -0.863
Hardness   0.388

                Factor1
SS loadings       2.710
Proportion Var    0.542

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 39.57 on 5 degrees of freedom.
The p-value is 1.82e-07
```

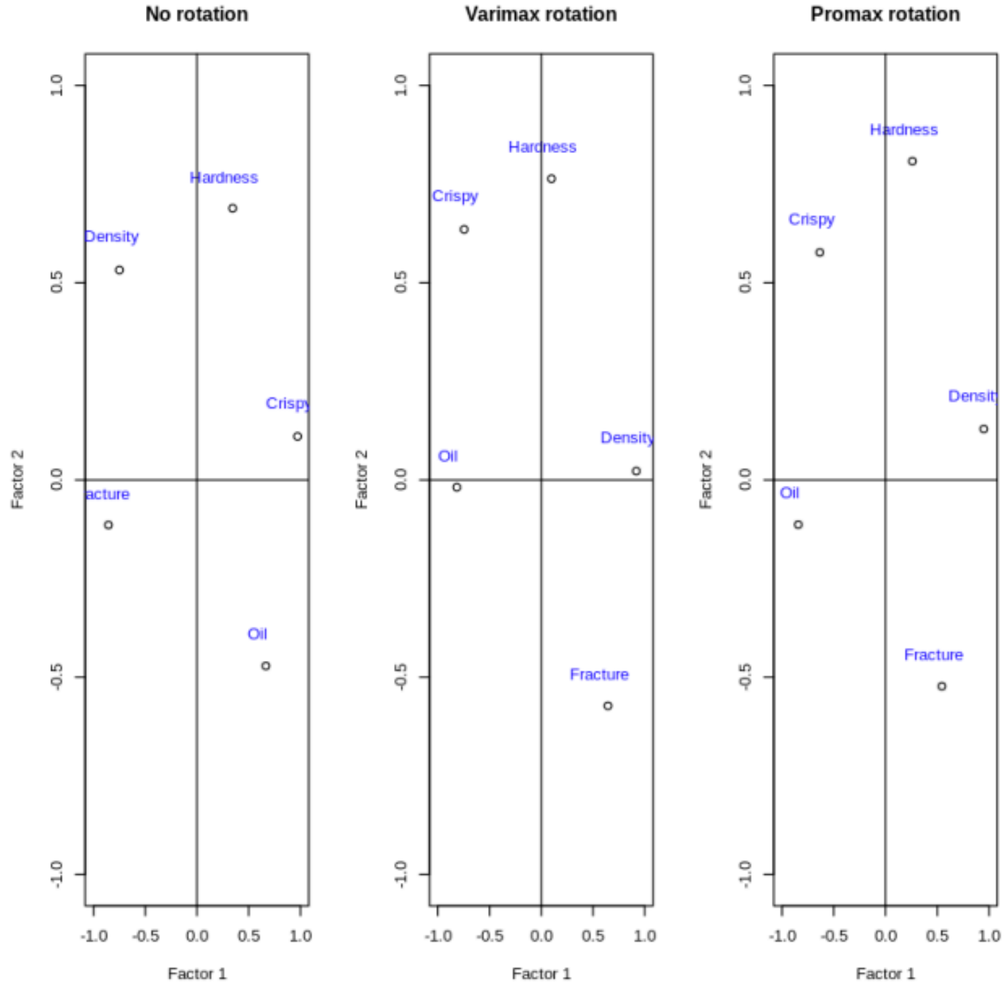
```
Call:
factanal(x = food, factors = 2)

Uniquenesses:
     Oil  Density   Crispy Fracture Hardness
   0.334    0.156    0.042    0.256    0.407

Loadings:
          Factor1 Factor2
Oil       -0.816
Density    0.919
Crispy    -0.745   0.635
Fracture   0.645  -0.573
Hardness           0.764

                Factor1 Factor2
SS loadings       2.490   1.316
Proportion Var    0.498   0.263
Cumulative Var    0.498   0.761

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.27 on 1 degree of freedom.
The p-value is 0.603
```



5. Ejemplo Práctico

laSalle
UNIVERSIDAD RAMON LLULL

Análisis con 1 Factor

Unicidad (Uniqueness):

- Representa la varianza específica de cada variable que no puede ser explicada por el factor.
- Variables como **Crispy (0.045)** y **Fracture (0.255)** están bien explicadas por el único factor, mientras que **Hardness (0.850)** no es explicada adecuadamente.

Cargas factoriales:

- El único factor parece estar relacionado principalmente con las variables **Crispy (0.977)** y **Fracture (-0.863)**.
- Las demás variables (e.g., Hardness y Oil) tienen menor influencia.

Conclusión:

- Un solo factor no parece suficiente para explicar todas las relaciones, ya que las variables como Hardness tienen alta unicidad (mayor ruido o varianza residual).
- La proporción de varianza explicada por este único factor es del **54.2%**, lo cual es moderado.

Análisis con 2 Factores

Unicidad (Uniqueness):

- Mejora en comparación con el modelo de un factor:
  - Las unicidades son menores para todas las variables.
  - La variable **Density** ahora tiene una unicidad de 0.156, lo que indica que está mejor explicada.

Cargas factoriales:

- **Factor 1:**
  - Explica principalmente **Density (0.919)** y **Hardness (0.764)**.
  - Representa propiedades físicas densas y resistentes.
- **Factor 2:**
  - Explica **Crispy (-0.745)** y **Fracture (0.645)**.
  - Representa propiedades relacionadas con fragilidad y crocancia.

Varianza explicada:

- El modelo de 2 factores explica un total de **76.1%**de la varianza:
  - Factor 1: 49.8%.
  - Factor 2: 26.3%.
- Esto es una mejora sustancial frente al modelo de un solo factor.

Conclusión:

- Dos factores son suficientes para capturar las relaciones subyacentes en los datos, ya que explican un porcentaje alto de la varianza y muestran cargas claras.

3. Gráficos de Rotaciones

Sin rotación:

- Las cargas factoriales están mezcladas, lo que dificulta la interpretación.
- Las variables no están claramente asociadas con un solo factor.

Rotación Varimax (Ortogonal):

- Mejora la interpretación:
  - **Hardness y Density** están claramente asociadas con el **Factor 1**.
  - **Crispy y Fracture** están claramente asociadas con el **Factor 2**.
  - La ortogonalidad asume que los factores no están correlacionados.

Rotación Promax (Oblicua):

- Permite correlación entre factores.
- Las asociaciones son similares a Varimax, pero los factores pueden estar correlacionados (más realista en datos complejos).