ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Técnicas no supervisadas: Análisis de Componentes Principales y Análisis Factorial

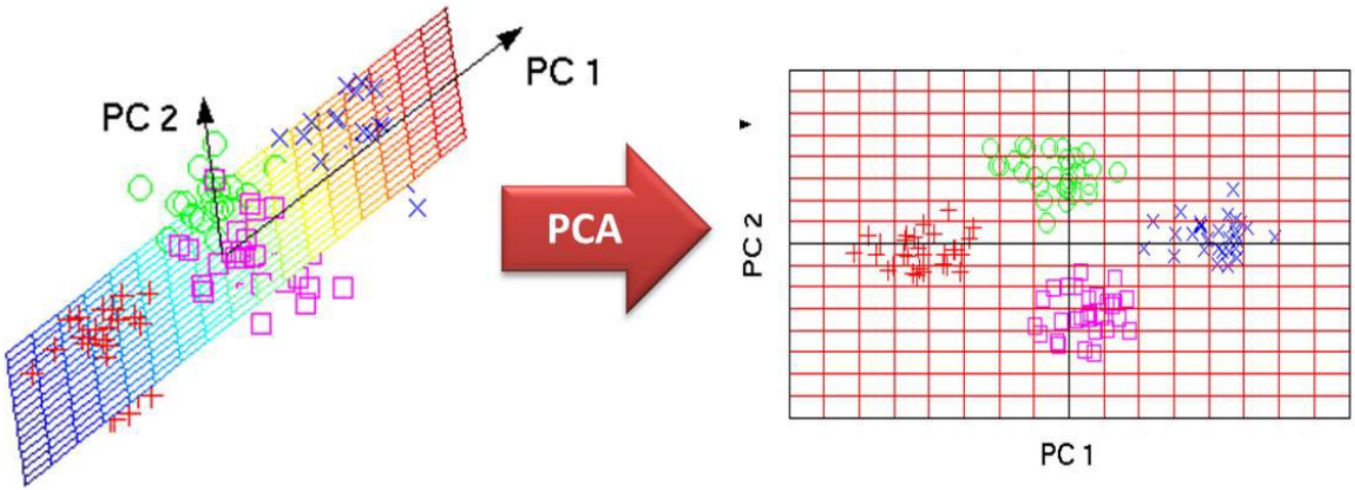
Minería de Datos y Modelización Predictiva

Máster Big Data, Data Science & Business Analytics
Universidad Complutense de Madrid
Curso 2024-2025

Entre las técnicas no supervisadas, se encuentran aquellas orientadas a la reducción de la dimensionalidad del conjunto de datos.

GENERALIDADES SOBRE LAS TÉCNICAS DE REDUCCIÓN DE DIMENSIONES

- **Objetivo**: Reducir el número de variables del conjunto
- **Tipos de datos**: Conjunto de datos con variables correlacionadas
- **Utilidades**:
  - Resumir infomación para su mejor interpretación
  - Evitar el problema de la relación entre predictores de cara a la modelización predictiva
  - Potencialmente puede eliminar ruido en los datos



Modelo matemático ACP

El objetivo principal es evaluar la posibilidad de representación de la información recogida en un conjunto de m variables, en un nuevo conjunto de p componentes **incorreladas** y extraídas mediante **combinación lineal** de las variables originales:

$$C_{1i} = a_{11}X_{1i} + a12X_{2i} + \ldots + a1mX_{mi} \quad i = 1, ..., p$$

Características ACP

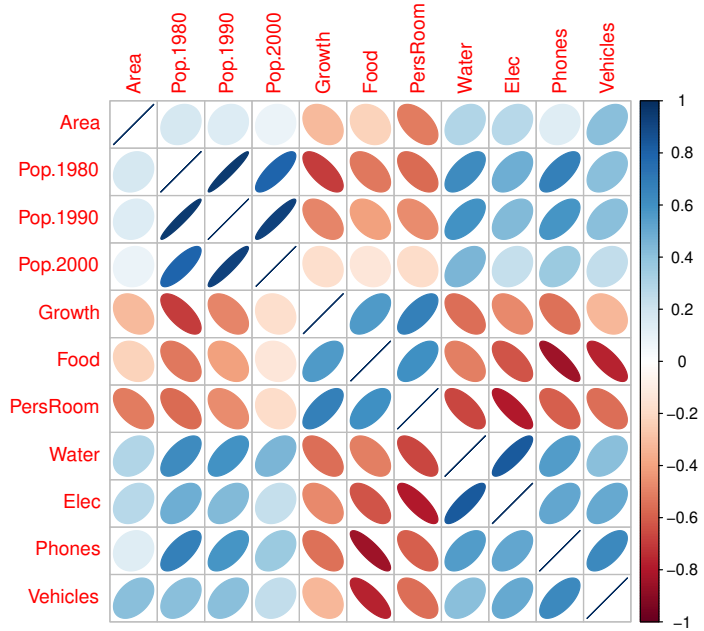
Para encontrar las mejores combinaciones lineales de las varables originales, se realiza la descomposición en valores singulares de la matriz de covarianza (datos homogéneos) o correlación (datos homogéneos) del conjunto original.

- Los **autovalores** o valores propios de la descomosición representan las **varianzas** de las nuevas componentes.
- Los **autovectores** o vectores propios representan las **direcciones** de los ejes de las nuevas componentes
- El primer autovalor es el que recoje mayor proporción de variabilidad de los datos, decreciendo sucesivamente a través de las componentes

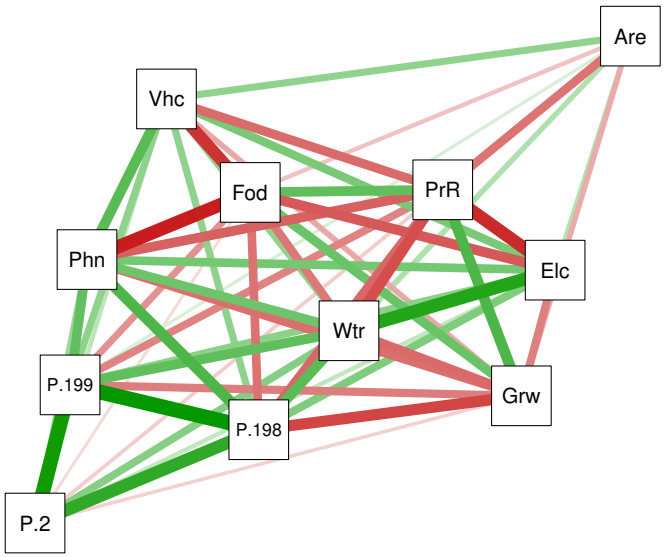
Criterios para la selección del número de componentes a retener

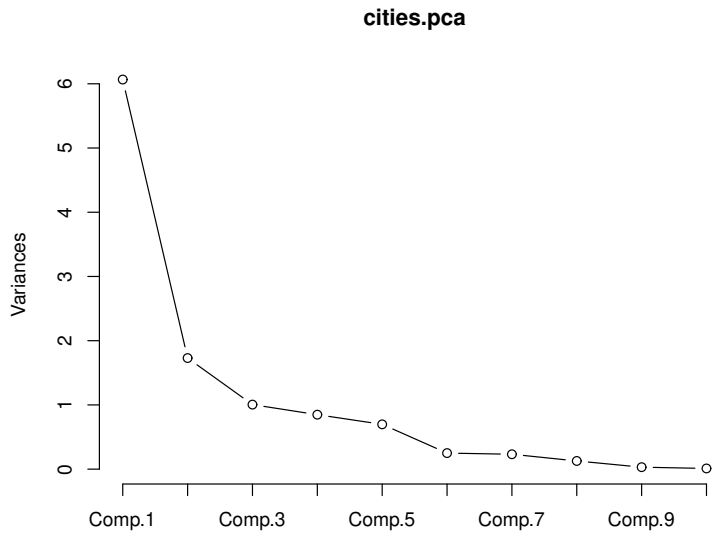
- **Criterio de varianza**. Retener un número de componentes que consiga explicar un mínimo de la variabilidad de los datos originales. Dificil determinar un umbral, depende del problema y de los objetivos del análisis (exploración vs. preparación para posterior predicción)
- **Critério de Kaiser**. Retener las componentes con varianzas/autovalores mayores que la varianza media esperada. En caso de análisis mediante matriz de correlaciones, autovalores mayores de la unidad.
- **Gráfico de sedimentación** (scree plot). Se grafican los autovalores y se selecciona el número de componentes previas a la sedimentación de la evolución de la varianza expicada por cada componente. Siempre decreciente, deseable una gran disminución en número bajo de componentes.
- **Dos componentes**. Solución interpretativa. Fácil de representar. En caso de valorar la interpretabilidad por encima de la varianza explicada.

Una solución ideal sería obtener una reducción a 2 dimensiones con una varianza explicada digamos mayor a un 80%. La realidad es, a menudo, bastante más compleja y deberemos valorar las soluciones obtenidas en el contexto del problema concreto.



	Area	Pop.1980	Pop.1990	Pop.2000	Growth	Food	PersRoom	Water	Elec	Phones	Vehicles
Area	1	0.18	0.15	0.08	−0.31	−0.23	−0.52	0.29	0.28	0.14	0.41
Pop.1980	0.18	1	0.96	0.79	−0.69	−0.52	−0.57	0.62	0.48	0.67	0.42
Pop.1990	0.15	0.96	1	0.92	−0.49	−0.41	−0.46	0.6	0.43	0.59	0.42
Pop.2000	0.08	0.79	0.92	1	−0.17	−0.14	−0.19	0.45	0.23	0.37	0.24
Growth	−0.31	−0.69	−0.49	−0.17	1	0.56	0.67	−0.56	−0.47	−0.54	−0.32
Food	−0.23	−0.52	−0.41	−0.14	0.56	1	0.6	−0.5	−0.62	−0.84	−0.76
PersRoom	−0.52	−0.57	−0.46	−0.19	0.67	0.6	1	−0.66	−0.79	−0.59	−0.55
Water	0.29	0.62	0.6	0.45	−0.56	−0.5	−0.66	1	0.83	0.55	0.42
Elec	0.28	0.48	0.43	0.23	−0.47	−0.62	−0.79	0.83	1	0.52	0.5
Phones	0.14	0.67	0.59	0.37	−0.54	−0.84	−0.59	0.55	0.52	1	0.63
Vehicles	0.41	0.42	0.42	0.24	−0.32	−0.76	−0.55	0.42	0.5	0.63	1





Comdiciones iniciales para ACP

El ACP se aplica sobre datos de naturaleza numérica. Una importante condición adicional deseable es la correlación inicial de las variables implicadas en el análisis, ya que de lo contrario no existiría información común que explicar en dimensiones reducidas.

- **Matriz de correlación**. Se inspecciona en busca de valores suficientemente altos en valor absoluto.
- **Matriz de p-valores de correlación**. Los p-valores deben ser próximos a cero, rechazando la hipótesis de correlación no significativa.
- **Determinante de la Matriz de Correlación**. Toma un valor próximo a $\mathbf{1}$ cuando no existe correlación y $\mathbf{0}$ cuando la correlación es máxima.
- **Test de esfericidad de Bartlett** . Basado en la matriz de correlación. Contrasta la hipótesis de que los datos estan dispuestos en un conjunto hiper-esférico, lo que conlleva una falta de correlación de los datos. Interesa rechazar dicha hipótisis. Buscamos p-valores bajos.
- **Índice KMO/MSA** de Kaiser-Meyer-Olkin. Es el índice de adecuación muestral, contrasta la correlación observada entre pares de variables y sus correspondientes correlaciones parciales (una vez eliminado el efecto de otras variables presentes). KMO<0.5 –>Inaceptable. KMO>0.7 –>Bueno.

Unva vez obtenidas las combinaciones lineales que definen los componentes principales, es posible interpretar los coeficientes (cargas o saturaciones) en relación a la influencia de cada variable en la construcción del cada componente. Saturaciones altas indican contribuciones altas.
En este paso es posible interpretar las componentes, y atrubuirles un significado genérico, a partir del conjuto de variables con mayor peso en su construcción.

GRÁFICO DE SATURACIONES

Muestra en un plano de dos componentes (por defecto las dos primeras) la posición de las variables originales.

- **Eje horizontal**: Representa la primera componente. Variables que están lejos a izquierda o derecha tienen mucho peso en la constucción del primer eje.
- **Eje vertical**: Representa la segunda componente. Variables que están lejos a arriba o abajo tienen mucho peso en la constucción del segundo eje.
- **Proximidad al origen**: Indicador de la poca relevancia de la variable en la construcción de ambos ejes. Probablemente, la información de estas variables no se recogerá en la solución con ese número de componentes.

GRÁFICO DE PUNTUACIONES

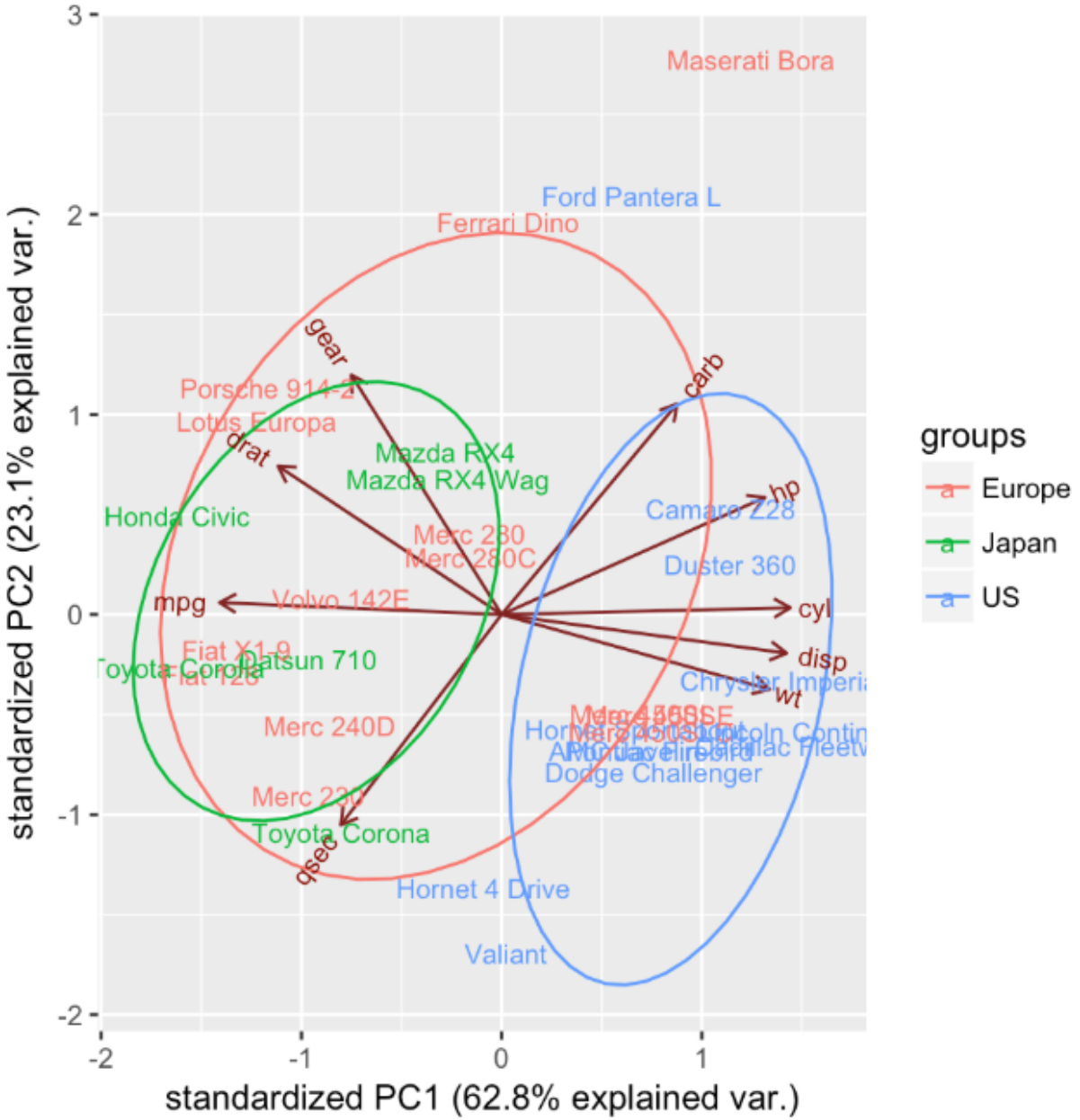
Muestra en un plano de dos componentes (por defecto las dos primeras) la posición de las observaciones del conjunto de datos. Es muy útil para:

- **Evaluar** grupos de registros con similares posiciones en el plano factorial identificando potenciales clusters.
- **Interpretar** al solución de componentes en términos de sugbrupos de datos formados por una eventual variable nominal del archivo

GRÁFICO CONJUNTO. BIPLOT

Muestra en un plano de dos componentes (por defecto las dos primeras) la posición de las observaciones y variables originales del conjunto de datos. En un sólo gráfico se evaluan las relaciones entre posibles entre variables originales, componentes y observaciones.

- Habitualmente se representan las variables con flechas desde el origen para dar una idea del ángulo (cos^2) que forman con las componentes.
  - Menor ángulo –> mayor correlación.
  - Mayor tamaño de flecha –> Mayor influencia en la construcción del eje.
- Habitualmente se representan las observaciones con puntos o etiquetas en el plano factorial.


standardized PC2 (23.1% explained var.)
standardized PC1 (62.8% explained var.)
3
2
1
0
-1
-2
-2
-1
0
1
Maserati Bora
Ford Pantera L
Ferrari Dino
Porsche 914-2
Lotus Europa
Mazda RX4
Mazda RX4 Wag
Honda Civic
Merc 230
Merc 280C
Volvo 142E
Fiat X1-9
Datsun 710
Merc 240D
Toyota Corona
Hornet 4 Drive
Valiant
Camaro Z28
Duster 360
Dodge Challenger
gear
carb
drat
hp
mpg
cyl
disp
wt
qsec
groups
Europe
Japan
US

MODELO MATEMÁTICO AF

El Análisis Factorial utiliza un **modelo inverso** al del Análisis de Componentes Principales:

- Las **variables** observables y medibles, aquellas de las que tenemos datos, se expresan como **combinación lineal** de un conjunto reducido de **factores latentes**, no observables, ni medibles directamente.
- Los **factores** comunes, generalmente pocos, son los que **explican todas las variables**.

Cada variable observable $\boldsymbol{X_i}$ es una combinación distinta de los mismos factores $\boldsymbol{F_j}$. Para completar el modelo se añade a cada variable $\boldsymbol{X_i}$ un **factor específico** $\boldsymbol{e_i}$, que recoge la **variabilidad** que no consigue explicar el conjunto de factores comunes; así, en un modelo con n variables observadas y dos factores comunes:

$$\boldsymbol{X_1 = \lambda_{11} F_1 + \lambda_{12} F_2 + e_1}$$
$$\boldsymbol{X_2 = \lambda_{21} F_1 + \lambda_{22} F_2 + e_2}$$
$$\boldsymbol{\dots}$$
$$\boldsymbol{X_n = \lambda_{n1} F_1 + \lambda_{n2} F_2 + e_n}$$

Se denomina **saturaciones** a los coeficientes del modelo, que deben ser estimados: $\boldsymbol{\lambda_{ij}}$ es la saturación de $\boldsymbol{X_i}$ en $\boldsymbol{F_j}$.

Estimación de los coeficientes

Existen varios procedimientos para la estimación de los coeficientes:

- Método de componentes principales (por esa razón a menudo se confunden ambas técnicas, que son conceptualmente distintas).
- Método de mínimos cuadrados no ponderados.
- Método de mínimos cuadrados generalizado.
- Método de máxima verosimilitud.
- Factorización de ejes principales.

Propiedades

A partir del modelo anterior se puede expresar la varianza de cada una de las variables X_i del siguiente modo:

$$\sigma_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \Phi_i \quad \Phi_i = Var(e_i)$$

- Es decir, la varianza de cada una de las variables es igual a la **suma de los cuadrados de las saturaciones**, que se denomina **comunalidad**, más un componente residual o **unicidad**.
- Ambas cantidades se expresan a menudo en **porcentaje de varianza** total de la variable.
- Cuanto más **altas** sean las **comunalidades** (parte de la varianza explicada por los factores comunes), mejor funciona el modelo factorial.
- Cuando el número de factores comunes es mayor que 1 el problema de la estimación de los coeficientes del modelo o saturaciones no está determinado: existen **infinitas soluciones**.
- Encontrada una solución, se puede obtener una transformación suya **rotación** que también es una solución válida.

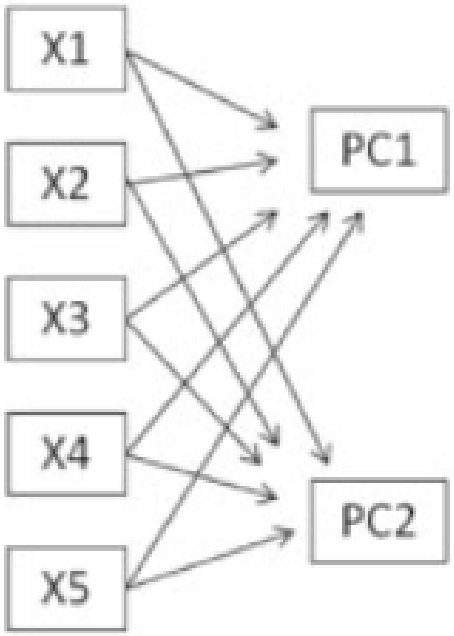
Tipos de Rotaciones

Existen varios procedimientos para la rotación de los factores con el objetivo de una mayor interpretabilidad:

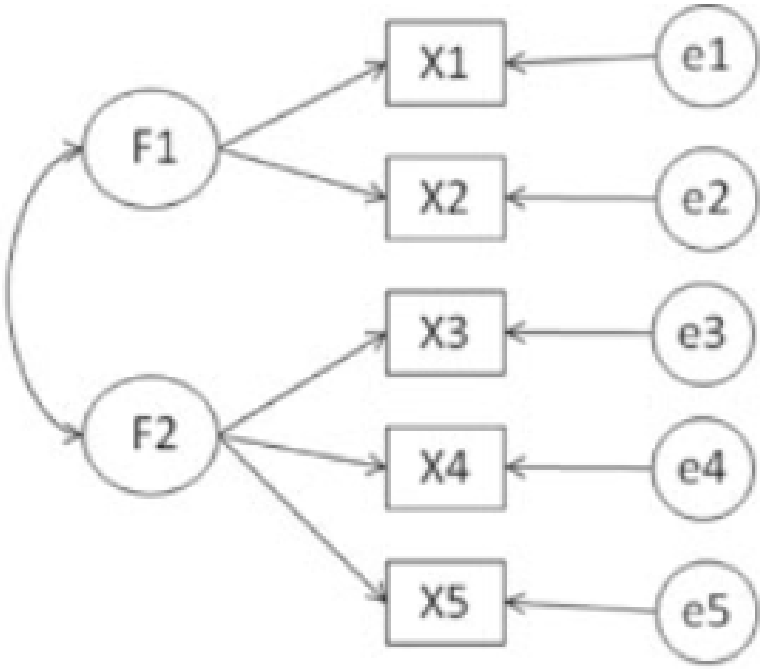
- **Rotaciones ortogonales**: Mantienen la independencia o **incorrelación de los factores**:
  - **Varimax**: máxima varianza de λ_{ij}^2 para cada F_i (simplifica los factores).
  - **Quartimax**: máxima suma de λ_{ij}^4 para cada x_j (simplifica las variables).
  - **Equamax**: intermedia (trata de simplificar factores y variables al mismo tiempo).
- **Rotaciones oblicuas** (los factores están correlacionados entre sí, dejan de ser independientes): entre ellas destaca **Promax**

- En la mayoría de las aplicaciones será adecuada la rotación varimax, ya que generalmente tendremos muchas variables y pocos factores; la expresión de cada variable será necesariamente sencilla, y solo necesitamos simplificar la expresión de los factores.
- En un modelo con muchos factores puede ser conveniente utilizar quartimax, para simplificar la expresión de cada variable.
- En algunos casos una rotación oblicua puede ser más satisfactoria, mejorando la interpretabilidad aunque con el sacrificio de la pérdida de independencia de los factores.

- En el **Análisis Factorial** cada variable es una función de los factores, en el **Análisis de Componentes Principales** cada componente es una función de las variables originales (los **modelos matemáticos son inversos**).
- Análisis de Componentes Principales tiene una única solución, Análisis Factorial infinitas (**rotaciones**).
- En el Análisis Factorial (confirmatorio) existe un modelo previo, indicando el número de factores y su significado. En el ACP no existe ninguna idea preconcebida similar.
- En el Análisis Factorial existen **múltiples métodos de estimación** de los coeficientes (Análisis de Componentes Principales es uno de ellos).
- El Análisis de Componentes Principales analiza (factoriza) la matriz de covarianzas o de correlaciones, el **Análisis Factorial** analiza la **matriz de correlaciones corregida**, sustituyendo los unos de la diagonal por la **comunalidad estimada**.



(a) Principal Components Model



(b) Factor Analysis Model

SIGUE SIENDO TEMA DE ESTUDIO...

Las aplicaciones de ACP y AF siguen estando vigentes hoy en día en muchos campos de investigación. Entre ellos destaca el campo de la genómica como ejemplo de aplicaciones en Big Data. Aún hoy se siguen comparando ambos métodos...

Principal Component Analysis and Factor Analysis: differences and similarities in Nutritional Epidemiology application

July 2019 · Revista Brasileira de Epidemiologia 22(2)

DOI: 10.1590/1980-549720190041

Benchmarking principal component analysis for large-scale single-cell RNA-sequencing

Koki Tsuyuzaki, Hiroyuki Sato, Kenta Sato & Itoshi Nikaido

Genome Biology **21**, Article number: 9 (2020) | Cite this article

Common Factor Analysis versus Principal Component Analysis: A Comparison of Loadings by Means of Simulations

January 2016 · Communication in Statistics- Simulation and Computation 45(1):299-321

DOI: 10.1080/03610918.2013.862274