TEMA 2

MÓDULO:
TÉCNICAS AVANZADAS DE DATA MINING

ANÁLISIS PCA Y FACTORIAL

FERRAN ARROYO

Licenciado en Empresariales y en Ciencias Actuariales y Financiaras por la UB. Máster Executive en Data Science por la MBIT School. Data Scientist.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Conocer la técnica del PCA.
- Conocer la técnica del Análisis Factorial.
- Entender la diferencia entre ambas técnicas.

1. INTRODUCCIÓN AL PCA

El **Análisis de Componentes Principales** (también denominado PCA) es un método, comúnmente, utilizado para reducir la dimensionalidad de conjuntos de datos con muchas variables.

El principal objetivo de este método consiste en sintetizar la información contenida en todas las variables de un conjunto de datos, transformándolas en otras nuevas variables (cuanto menos mejor), siempre intentando mantener la mayor parte de la información del conjunto de variables inicial.

Cabe destacar que la utilización de este método implica, automáticamente, una pérdida de información sobre nuestro dataset original a cambio de una reducción considerable en el número de variables.

Así pues, **la idea de PCA es simple: reducir el número de variables de un conjunto de datos, mientras se preserva la mayor cantidad de información posible.**

Este método te será muy útil cuando tengas que reducir y condensar la información de cualquier conjunto de datos inmenso.

CITA

"Too much of anything is good for nothing."

Anónimo

Sin embargo, no siempre será necesario utilizarlo cuando tengas conjuntos de datos con muchas variables.

CITA

"Too much of anything is bad but too much good whiskey is barely enough."

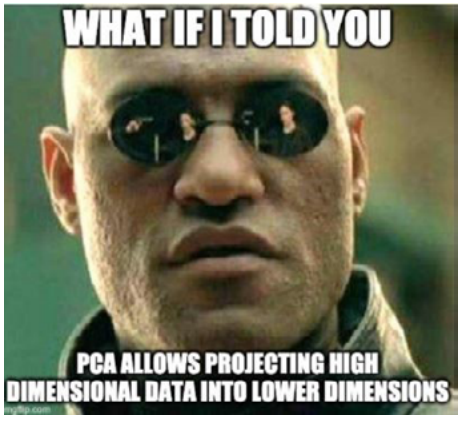
Mark Twain

Vamos a ilustrar este problema con un pequeño ejemplo:

Imaginemos que tenemos un conjunto de datos de dimensión 300 (n) × 50 (p) donde **n** representa el número de observaciones y p representa el número de predictores.

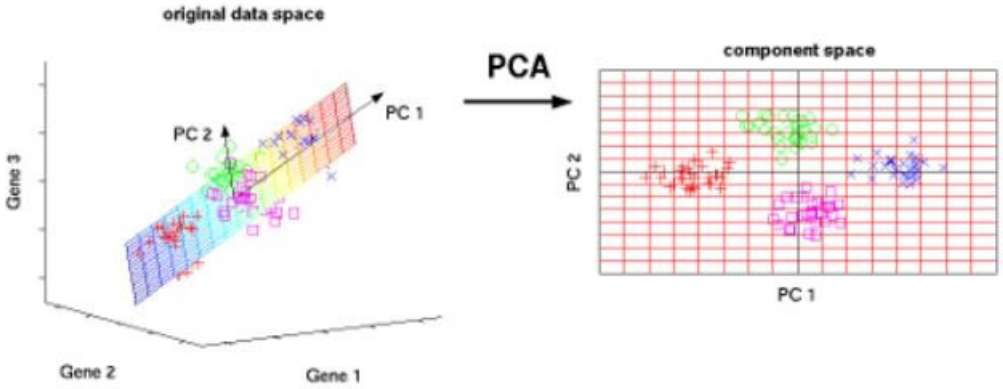
Dado que tenemos una gran p = 50, puede haber p(p-1)/2 gráficos de dispersión, es decir, más de 1000 gráficos posibles para analizar la relación de las variables.

¿No sería un trabajo sumamente pesado realizar un análisis exploratorio de estos datos?



Fuente: https://towardsdatascience.com/pca-principal-component-analysis-explained-visually-in-5-minutes-20ce8a9ebf0f

En la siguiente imagen tienes un ejemplo visual de cómo este método puede transformar un conjunto de datos de 3 dimensiones en otro nuevo conjunto de datos con 2 dimensiones mediante la utilización del PCA. Es importante reseñar que cada una de las nuevas dimensiones es una combinación lineal entre las dimensiones iniciales.



Fuente: http://www.nlpca.org/pca_principal_component_analysis.html

1.1 QUÉ ES UN COMPONENTE PRINCIPAL

Un componente principal **es una combinación lineal normalizada de las variables originales en un conjunto de datos.** En la imagen de arriba, PC1 y PC2 son los componentes principales.

Imaginemos que tenemos un conjunto de variables: X^1, X^2...Xp.

El componente principal puede definirse como:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi p^1 Xp$$

donde,

- Z^1 es el primer componente principal.
- Φp^1 es el vector de carga que comprende las cargas (Φ^1, Φ^2..) del primer componente principal. Las cargas están limitadas a una suma de cuadrados igual a 1. Esto se debe a que la gran magnitud de las cargas puede dar lugar a una gran variabilidad. También, se define la dirección del componente principal (Z^1) a lo largo de la cual los datos varían más. Da como resultado una línea en el espacio dimensional p que es la más cercana a las n observaciones. La cercanía se mide utilizando la distancia euclídea cuadrada media.
- X^1..Xp son las variables estandarizadas. Las variables estandarizadas tienen una media igual a cero y una desviación estándar igual a uno.

Por lo tanto,

El primer componente principal es una combinación lineal de variables originales que capta la máxima

varianza del conjunto de datos, determinando la dirección de la mayor variabilidad (varianza) en los datos.

Cuanto mayor sea la variabilidad capturada en el primer componente, mayor será la información capturada por el componente.

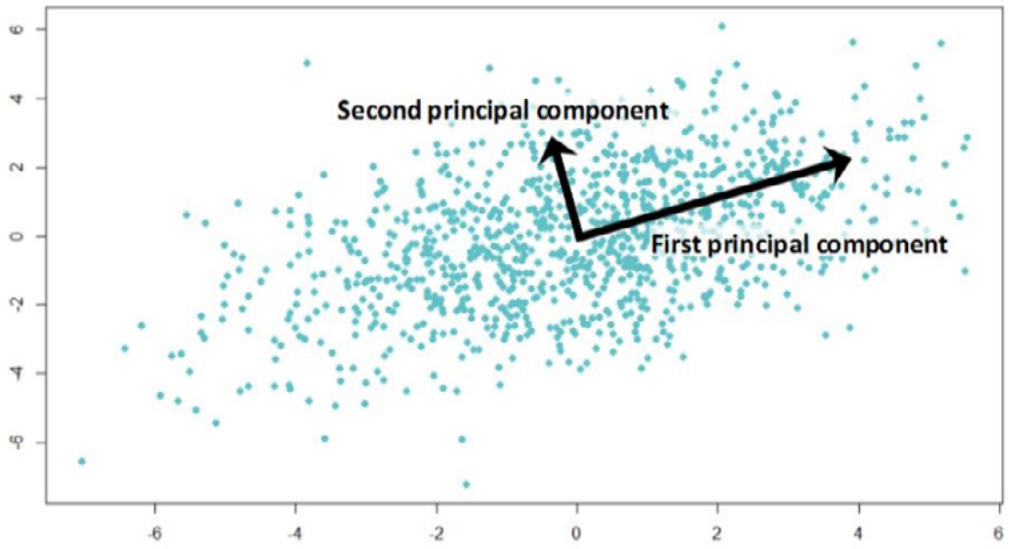
Ningún otro componente puede captar una variabilidad mayor que el primer componente principal. De manera similar, también podemos calcular el segundo componente principal.

El segundo componente principal (Z^2) es también una combinación lineal de variables originales que capta la varianza restante en el conjunto de datos y no está correlacionada con Z^1. En otras palabras, **la correlación entre el primer y el segundo componente debería ser cero.** Se puede representar como:

$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi p2Xp$$

Si los dos componentes no están correlacionados, sus direcciones deben ser ortogonales (ver imagen inferior).

Esta imagen se basa en una simulación de datos con 2 componentes. Fíjate en la dirección de los componentes, se espera que sean ortogonales. Esto sugiere la correlación nula o casi nula entre los componentes:



Fuente: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

1.2 ESTANDARIZACIÓN DE LAS VARIABLES

Los componentes principales se han de calcular con la versión estandarizada de las variables originales. Esto es debido a que, habitualmente, estas variables tienen distintas escalas.

EJEMPLO

Imaginate un conjunto de datos con unidades de medida de variables como litros, kilómetros, años, etc. Está claro que la escala de varianzas de estas variables será grande.

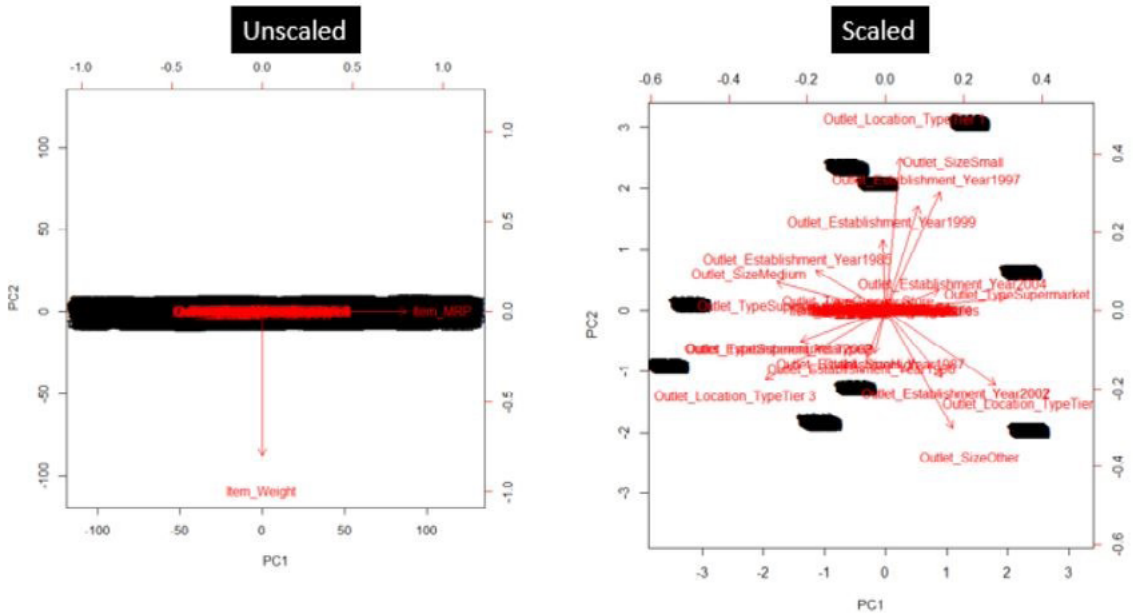
Realizar PCA en variables no normalizadas conllevará una dependencia del componente principal a las variables que tengan una varianza más elevada, y esto no es nada deseable.

Fuente: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

Matemáticamente, realizar el proceso de estandarización es muy sencillo:

$$z = \frac{value - mean}{standard\ deviation}$$

En la siguiente imagen puedes ver, gráficamente, un PCA calculado sobre un conjunto de datos, dos veces (con variables no escaladas y con variables escaladas). Este conjunto de datos tiene ~40 variables. Como puedes ver, el primer componente principal está dominado por la variable Item_MRP, y el segundo componente principal está dominado por la variable Item_Weight. Esta dominación prevalece debido al alto valor de la varianza asociada a una variable. Cuando las variables se escalan, obtenemos una representación mucho mejor de las variables en el espacio 2D:

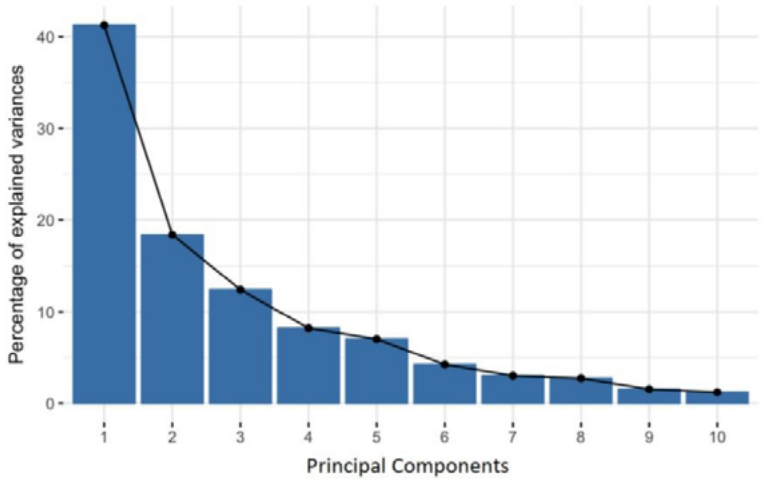


Fuente: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

1.3 VECTORES PROPIOS Y VALORES PROPIOS

Los vectores propios y los valores propios son los conceptos de álgebra lineal que necesitamos computar a partir de la matriz de covarianza para determinar los componentes principales. Antes de llegar a la explicación de estos conceptos, entendamos primero qué entendemos por componentes principales.

Los componentes principales son nuevas variables que se construyen como combinaciones lineales o mezclas de las variables iniciales. Estas combinaciones se hacen de tal manera que las nuevas variables (es decir, los componentes principales) no están correlacionadas y la mayor parte de la información de las variables iniciales se comprime en los primeros componentes. Así pues, la idea es que los datos de 10 dimensiones te dan 10 componentes principales, pero **el PCA intentará poner el máximo de información posible en el primer componente, luego el máximo de información restante en el segundo y así, sucesivamente,** hasta tener algo parecido a lo que se muestra en el siguiente gráfico:



Fuente: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Organizar la información en componentes principales de esta manera, **te permitirá reducir la dimensionalidad sin perder mucha información,** ya que podrás descartar los componentes con poca información y considerar los componentes restantes como las nuevas variables.

IMPORTANTE

Una cosa importante a tener en cuenta aquí, es que los componentes principales son poco interpretables y no tienen ningún significado real ya que están construidos como combinaciones lineales de las variables iniciales.

Geométricamente hablando, los componentes principales representan las direcciones de los datos que explican una cantidad máxima de varianza, es decir, las líneas que capturan la mayor parte de la información de los datos. **La relación entre la varianza y la información aquí, es que, cuanto mayor es la varianza transportada por una línea, mayor es la dispersión de los puntos de datos a lo largo de ella, y cuanto mayor es la dispersión a lo largo de una línea, más información tiene.** Para poner todo esto de forma sencilla, sólo hay que pensar en los componentes principales como nuevos ejes que proporcionan el mejor ángulo para ver y evaluar los datos, de modo que las diferencias entre las observaciones sean más visibles.

1.4 CONSTRUCCIÓN DE LOS COMPONENTES PRINCIPALES

A la hora de trabajar con componentes principales, debes tener en cuenta que habrá tantos como variables existan en tu conjunto de datos. Estos se construyen de tal modo que el primer componente siempre es el que aglutinará la mayor varianza posible. En otras palabras, el primer componente debe ser el que contenga más explicabilidad de la información que el resto.

Otro punto importante que debes tener en cuenta, es que los vectores propios y los valores propios siempre vienen en parejas, por lo que cada vector propio tiene un valor propio asociado.

EJEMPLO

Un conjunto de datos con 3 variables tendrá 3 vectores propios con sus 3 valores propios asociados.

Así pues, ya habrás podido deducir que la magia que se esconde detrás de la afirmación de Morfeo viene dada por estos dos elementos. Los valores contenidos en los vectores propios no son más que las direcciones de los ejes donde hay más varianza, mientras que los valores propios son coeficientes adjuntos a los vectores propios que contienen la información acerca de la cantidad de varianza transportada en cada componente principal.

Vamos a suponer que tenemos un conjunto de datos con 2 variables x,y cuyos vectores propios (eigenvectors) y valores propios (eigenvalues) son los siguientes:

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

Si rankeamos los valores propios en orden descendente, podemos deducir que λ1>λ2, lo que significa que el vector propio v1 se corresponde con el primer componente principal (PC1), mientras que el vector propio v2 se corresponde con el segundo componente principal (PC2).

Después de obtener cada uno de los componentes principales, para calcular el porcentaje de varianza (información) que representa cada componente, dividimos el valor propio de cada componente por la suma de los valores propios.

Aplicándolo a este ejemplo:

Varianza PC1 = λ1/(λ1+λ2) –> Representa el 96% de la varianza

Varianza PC2 = λ2/(λ1+λ2) –> Representa el 4% de la varianza

2. INTRODUCCIÓN AL ANÁLISIS FACTORIAL

En la vida real, los datos tienden a seguir algunos patrones. Algunos pueden ser aparentes, pero otros no tienen por qué serlo.

El propósito esencial del Análisis Factorial es describir las relaciones de covarianza entre varias variables en términos de unos pocos componentes aleatorios subyacentes e inobservables llamados factores.

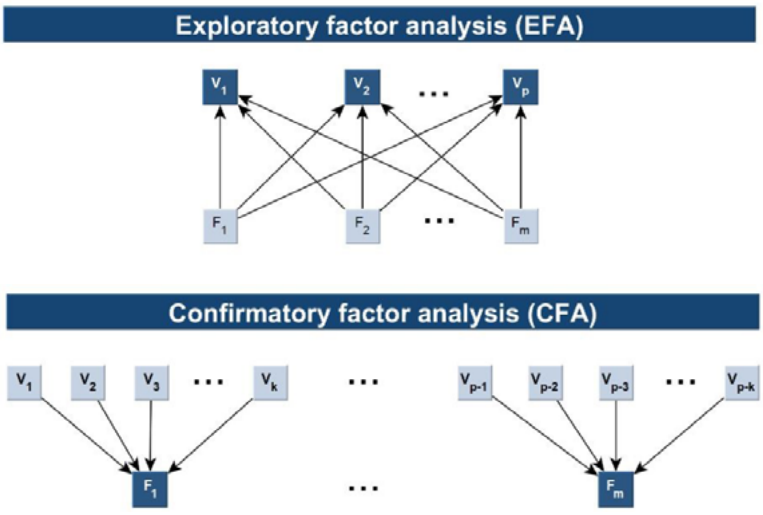
Las variables observadas se modelan como combinaciones lineales de factores más expresiones de error.

El análisis factorial se originó en psicometría, y se usa en las ciencias del comportamiento tales como ciencias sociales, marketing, gestión de productos, investigación operativa, y otras ciencias aplicadas que tratan con grandes cantidades de datos.

Existen dos tipos de Análisis Factorial:

- **Análisis Factorial Exploratorio (AFE):** se usa para tratar de descubrir la estructura interna de un número relativamente grande de variables. La hipótesis a priori del investigador es que pueden existir una serie de factores asociados a grupos de variables. Las cargas de los distintos factores se utilizan para intuir la relación de estos con las distintas variables. Este es el tipo de análisis factorial más común.
- **Análisis Factorial Confirmatorio (AFC):** Trata de determinar si el número de factores obtenidos y sus cargas se corresponden con los que cabría esperar a la luz de una teoría previa acerca de los datos. La hipótesis, a priori, es que existen unos determinados factores preestablecidos y que cada uno de ellos está asociado con un determinado subconjunto de las variables. El análisis factorial confirmatorio, entonces, arroja un nivel de confianza para poder aceptar o rechazar dicha hipótesis. También, considera las variables como dos medidas que pueden ser cuantificadas constantemente.

Como una imagen vale más que mil palabras:



Fuente: https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/index.html

Para este tipo de análisis, es clave asumir que **las variables pueden ser agrupadas mirando sus correlaciones.** Es decir, asumiremos que todas las variables de un grupo específico tienen una **alta correlación entre ellas, pero una baja correlación con las variables de otros grupos.** En ese caso, podemos pensar en cada grupo de variables como una representación de una única construcción subyacente, o un factor que es responsable de la correlación observada.

Por ejemplo, la correlación de un grupo de variables compuesto por las puntuaciones obtenidas por un estudiante en matemáticas, biología y física podrían provenir de un "factor de inteligencia" subyacente mientras que otro grupo de variables compuesto por las puntuaciones obtenidas en educación física podría provenir de otro tipo de factor subyacente.

2.1 EL MODELO FACTORIAL (EFM)

Consideramos un conjunto de variables definidas como x1,x2,...,xp, y el promedio de esas variables definidas como μ1,μ2,...,μ, así como su matriz de covarianzas definida como Σ:

$$\mathbf{X}_{p\times 1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \qquad \mu_{p\times 1} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}, \qquad \mathbf{\Sigma}_{p\times p} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix}$$

Para los cuáles consideraremos m factores, definidos como F1,F2,...,Fm:

$$\mathbf{F}_{m\times 1} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

La idea básica detrás del Análisis Factorial es la de una regresión, lo que significa que expresamos cada una de las variables observadas como combinaciones lineales de variables (o factores) latentes. Así, si tenemos variables manifiestas p y factores m:

$$x_j = \mu_j + \lambda_{j1}F_{j1} + \lambda_{j2}F_{j2} + \ldots + \lambda_{jm}F_{jm} + e_j \qquad j = 1, 2, \ldots, p,$$

donde asumimos que los factores F1,F2,...,Fm tienen media y desviación estándar igual a 0. El termino de error ej también se asume con media y desviación estándar igual a 0, σj.

Expresado en matrices resultaría en:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pp} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix}$$

Lo cuál puede ser expresado con notación matricial:

$$\mathbf{X}_{p\times 1} = \mu_{p\times 1} + \Lambda_{p\times m} F_{m\times 1} + \mathbf{e}_{p\times 1}$$

Reescribiéndolo, obtenemos el Modelo Factorial (EFM)

$$\mathbf{X} - \mu = \Lambda F + \mathbf{e}$$

2.2 LA VARIANZA EXPLICADA

Una vez tenemos nuestro modelo especificado, estamos interesados en saber si funciona bien o no, o en otras palabras, qué cantidad de variabilidad en X, dada la matriz de covarianzas Σ donde:

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T$$

viene explicada por el Modelo Factorial.

Supongamos que existen p variables originales, para ser explicadas con m factores (m < p). **El Análisis Factorial descompone la matriz de covarianza**s definida como Σ de las variables originales X en una matriz de cargas de dimensión pxm, definida como Λ, donde Λ=Cov(XF), y una matriz de varianza no explicada, definida como Ψ, donde Ψ=Cov(e), tal que:

$$\Sigma = \Lambda\Lambda^T + \Psi$$

Esta ecuación indica que conocemos la variabilidad en X, definida por Σ, si conocemos la matriz de cargas Λ y la diagonal de la matriz de las varianzas no explicadas Ψ.

Así, podemos explicar, conceptualmente, Σ con dos términos. El primer término, la matriz de cargas Λ, da los coeficientes (λjm) que relacionan los factores (Fjm) con cada observación particular (xj). Estos coeficientes pueden ser estimados a partir de los datos de las observaciones. Por consiguiente, el término ΛΛT corresponde a **la variabilidad, que puede ser explicada por los factores.** Esta proporción de la variabilidad global, explicada como una combinación de factores, se denomina comunalidad. En cambio, **la proporción de la variabilidad, que no puede explicarse por una combinación lineal de los factores,** dada por el término Ψ se denomina singularidad:

$$\Sigma = \underbrace{\Lambda\Lambda^T}_{\text{communality}} + \underbrace{\Psi}_{\text{uniqueness}}$$

2.3 ESTIMACIÓN DE LOS PARÁMETROS

En la práctica, no conocemos Σ, pero podemos estimarlo mediante la matriz de covarianzas de la muestra S (o mediante la matriz de correlación de la muestra R, si queremos tratar con variables normalizadas). Por lo tanto, buscamos las estimaciones Λ^ y Ψ^ de manera que ΛΛT+Ψ se acerque lo máximo posible a S. En otras palabras, intentamos que la discrepancia entre la matriz de covarianza de la muestra, S y la

matriz modelo Σ sea lo más pequeña posible.

El método de estimación implementado en la función R factanal() es el método de máxima verosimilitud. Las estimaciones por máxima verosimilitud se obtienen por iteración, un proceso en el que Λ^ y Ψ^ se alteran, sistemáticamente, para hacer que la función de máxima verosimilitud sea cada vez más pequeña. Este enfoque supone que el vector x se distribuye de acuerdo con una distribución normal multivariante, es decir:

$$f(x) = |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

La función de verosimilitud para una muestra de n unidades es:

$$\mathcal{L}(x, \mu, \Sigma) = |2\pi\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)}$$

Mientras que la función de verosimilitud logarítmica es:

$$\ln L(x, \mu, \Sigma) = -\frac{n}{2}\ln|2\pi\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)$$

Si el modelo de Análisis fectorial se mantiene Σ=ΛΛT+Ψ, la función de verosimilitud logarítmica:

$$\ln L(x, \Lambda, \Psi) = -\frac{n}{2}\ln|2\pi(\Lambda\Lambda^T+\Psi)| - \frac{n}{2}tr(\Lambda\Lambda^T+\Psi)^{-1}S$$

Por último, la función de verosimilitud debe maximizarse con respecto a Λ y Ψ. En los casos en que uno o más elementos de la estimación de Ψ se vuelven negativos, conocidos como el caso de Heywood, los elementos de la estimación de Ψ se limitan a ser no negativos.

RECUERDA

Una ventaja importante de ajustar el modelo factorial por máxima verosimilitud es que podemos evaluar el número de factores a incluir en el modelo aplicando una prueba de razón de probabilidad, que sigue una distribución de χ2 con 12[(p-m2-(p+m))] grados de libertad.

2.4 ROTACIÓN DE LOS FACTORES

Es importante darse cuenta de que el modelo de factor lineal no es identificable. Esto significa que hay dos o más parametrizaciones que son equivalentes desde el punto de vista de la observación o, en otras palabras, que existe un número infinito de matrices diferentes Λ que pueden generar los mismos valores x. Por eso, el análisis factorial se suele desarrollar en **dos etapas:**

- **En la primera,** se calcula un conjunto de cargas, Λ, que produce varianzas y covarianzas teóricas que se ajustan lo más posible a las observadas. Estas cargas, sin embargo, pueden no proporcionar una interpretación razonable.

- **En la segunda** etapa, las cargas, Λ se transforman con el objetivo de llegar a otro conjunto que se ajuste, igualmente, a las varianzas y covarianzas observadas, pero que sea más fácil de interpretar.

El proceso de transformación de un patrón de factores se denomina, generalmente, **rotación.** Hay dos tipos básicos de transformaciones:

- **Ortogonales:** mediante la rotación ortogonal se preserva la independencia de los factores. Es decir, los factores no están correlacionados.
- **Oblicuos:** mediante la rotación oblicua no se preserva la independencia de los factores. Es decir, los factores pueden estar correlacionados.

Dos métodos populares ampliamente utilizados son:

- **El método varimax:** realiza la transformación ortogonal. Maximiza la varianza de las cargas al cuadrado para cada factor, haciendo así que algunas de estas cargas sean lo más grandes posibles y el resto lo más pequeñas posibles en valor absoluto. En consecuencia, las variables se dividen en grupos, de manera que las cargas dentro de cada grupo son altas en un solo factor, moderadas a bajas en unos pocos factores y despreciables en los restantes factores. Este método fomenta la detección de factores cada uno de los cuáles esté relacionado por unas pocas variables y desalienta la detección de factores que puedan ser influenciados por muchas variables. Este es el método más común.
- **El método promax:** realiza la transformación oblicua.

PIENSA UN MINUTO

Entonces...¿El PCA y FA son lo mismo?

A veces hay confusión entre el análisis de componentes principales (PCA) y el análisis de factores (FA). **Ambos métodos tienen por objeto reducir la dimensionalidad de un vector de variables aleatorias, sin embargo, la diferencia fundamental es que el análisis factorial especifica, explícitamente, un modelo que relaciona las variables observadas con un conjunto más pequeño de factores subyacentes no observables, mientras que el PCA no.**

Este modelo puede ajustarse a los datos o no. Por el contrario, el PCA es sólo un método de transformación de las variables.

2.5 IMPLEMENTACIÓN EN R

Toda esta teoría está muy bien, pero ¿de qué puede servirnos?

Trabajar con datos multidimensionales puede ser un auténtico dolor de muelas, así que si podemos reducir el número de dimensiones, podemos hacer que los datos sean más fáciles de trabajar y tratar de interpretar, subjetivamente los factores subyacentes.

Existen distintas implementaciones del Análisis Factorial en R. La que aprenderás en este curso es la función **factanal()** incluida en el paquete base.

Los argumentos básicos de la función son:

- El argumento **"covmat"** debe ser una matriz de covarianzas o un dataframe con todas las observaciones y este será el input del proceso.
- El argumento **"factors"** hace referencia al número de factores a ajustar.
- El argumento **"scores"** nos permitirá realizar el cálculo de las puntuaciones mediante el estimador de Thompson o el de Bartlett. Para este argumento, podrás escoger los valores "regression" o "Barlett".
- El argumento **"rotation"** hace referencia al tipo de rotación utilizada. Podrás elegir los valores "varimax" (para la rotación ortogonal), "promax" (para la rotación obliqua), o "none" (si no quieres aplicar rotación).

3. ACTIVIDAD GUIADA

El conjunto de datos que utilizaremos para este ejemplo contiene información acerca de las características de distintas pastas. Las dimensiones son 50 filas (observaciones) y 5 columnas (variables/dimensiones). Las variables son:

- **Aceite:** porcentaje de aceite en la pasta.
- **Densidad:** la densidad del producto (cuanto más alto es el número, más denso es el producto).
- **Crujiente:** una medida que define la textura del alimento cuando al masticarlo cruje. Esta variable tiene una escala de 7 a 15, siendo 15 más crujiente.
- **Fractura:** el ángulo, en grados, a través del cual la pasta puede ser doblada, lentamente, antes de que se fracture.
- **Dureza:** se utiliza una punta afilada para medir la cantidad de fuerza necesaria antes de que se produzca la rotura.

Fuente: http://openmv.net/info/food-texture

3.1 CÁLCULO DEL ANÁLISIS FACTORIAL

```
# Lectura del dataset
food <- read.csv("https://userpage.fu-berlin.de/soga/300/30100_data_sets/food-texture.csv",
                 row.names = "X")
str(food)
```

```
'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 ...
```

Tal y como se puede observar, todas las variables de nuestro dataset son numéricas.

Además del conjunto de datos, la función factanal() requiere una estimación del número de factores para poder inicializarse. Este es un aspecto a tener en muy cuenta en el Análisis Factorial.

Si tenemos una hipótesis sobre las variables latentes, podemos empezar con nuestra suposición. Si no tenemos ninguna pista sobre el número de factores y el número de variables del conjunto de datos no es demasiado grande, se puede, simplemente, probar varios valores para inicializar el modelo. Otro enfoque más sofisticado consiste en utilizar el análisis de componentes principales para obtener una buena estimación inicial del número de factores.

En este ejemplo supondremos que hay 2 factores latentes y los fijaremos como valor.

Además, mantendremos los valores predeterminados para los parámetros de la función scores (score="-none") y rotation (rotation="varimax"):

```
# Cálculo del Análisis de Factores
food.fa <- factanal(food, factors = 2)
print(food.fa)
```

```
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
```

3.2 INTERPRETACIÓN DE RESULTADOS

Antes de interpretar los resultados del análisis factorial, recordemos la idea básica que hay detrás. El análisis factorial crea una línea combinaciones de factores para abstraer la comunalidad subyacente de las variables.

En la medida en que las variables tienen una determinada comunalidad, un menor número de factores son capaces de capturar la mayor parte de la varianza en el conjunto de datos. Esto nos permite agregar un gran número de variables observables en un modelo para representar un concepto subyacente, facilitando la comprensión de los datos.

La variabilidad en nuestros datos, X, vienen dados por Σ, cuyo estimador Σ^ está compuesto por la varia-

bilidad explicada como la combinación lineal entre los factores (comunalidad) más la singularidad:

$$\Sigma = \underbrace{\Lambda\Lambda^T}_{\text{communality}} + \underbrace{\Psi}_{\text{uniqueness}}$$

Dicho esto, pasamos a comentar los resultados de la función:

- El **primer output** es un simple recordatorio de los parámetros que hemos utilizado en la función.

- El **segundo output** es un vector en el que podemos ver **la singularidad de cada variable, cuyo rango va de 0 a 1. La singularidad, a veces definida como el ruido,** corresponde a la proporción de la variabilidad que no puede ser explicada como una combinación lineal entre los factores. Este valor se corresponde con el valor de Ψ visto en la ecuación anterior. **Una alta singularidad para una variable indica que los factores no representan bien su varianza.** Para llamar directamente a este output, deberás hacerlo del siguiente modo:

```
{r}
# Singularidad de cada variable
food.fa$uniquenesses
```

```
      Oil   Density    Crispy  Fracture  Hardness
0.3338599 0.1555255 0.0422238 0.2560235 0.4069459
```

- Para obtener las comunalidades, únicamente deberás restar 1 a cada una de las singularidades:

```
# Comunalidades
1- food.fa$uniquenesses
```

```
      Oil   Density    Crispy  Fracture  Hardness
0.6661401 0.8444745 0.9577762 0.7439765 0.5930541
```

- **El tercer output** es una tabla en la que podemos observar **la contribución de cada variable a cada uno de los factores. El rango que pueden tomar va de -1 a 1.** Las variables con valores elevados están bien explicadas por los factores. Ten en cuenta que puede haber variables que no tengan valores en algún factor, esto es debido a que R no imprime salidas con valores menores a 0.1. Si estás interesado en modificar este punto de corte, tienes más información tipeando help (loadings) en la consola.

Fuente: https://makeameme.org/meme/approved-by-schumacher

- **El cuarto output** muestra la **proporción de la varianza explicada por cada factor, así como la puntuación de cada uno de ellos.** En la fila "Cumulative Var", se muestra **la proporción acumulativa de la varianza explicada en rango 0-1**. En la fila "Proportion Var", se muestra la proporción de la varianza explicada por cada factor, individualmente. En la fila "SS loadings" se muestra la suma de cargas al cuadrado. Estos valores se suelen uti-

lizar para determinar el valor de un factor en particular. Suele merecer la pena mantener un factor si su valor es mayor a 1. (regla de Kaiser).

- El quinto y último output muestra los resultados de un contraste de hipótesis, cuya hipótesis nula (H0) es que el número de factores en el modelo (en nuestro ejemplo 2 factores) es suficiente para capturar la dimensionalidad completa del conjunto de datos. Convencionalmente y, como ya sabes, rechazamos H0 si el valor p es inferior a 0,05. Tal resultado indica que el número de factores es demasiado pequeño. Por el contrario, no rechazamos H0 si el valor p es superior a 0,05. Tal resultado indica que es probable que haya suficientes (o más que suficientes) factores que capturen la dimensionalidad completa del conjunto de datos. El alto valor p en nuestro ejemplo anterior nos lleva a no rechazar H0, e indica que hemos ajustado un modelo apropiado.

RECUERDA

Si el input es una una matriz de covarianzas y no un dataframe, el contraste de hipótesis no se proporcionará a menos que proporcionamos, explícitamente, el número de observaciones en el argumento "n.obs" como argumento adicional a la llamada de la función.

3.3 MATRIZ DE RESIDUOS

Si recuerdas la ecuación vista anteriormente:

$$\Sigma = \underbrace{\Lambda\Lambda^T}_{\text{communality}} + \underbrace{\Psi}_{\text{uniqueness}}$$

Utilizando nuestro modelo definido, anteriormente, como food.fa podríamos calcular Σ^ y compararlo con la matriz de correlaciones observada S mediante algebra matricial. A tal efecto, necesitaremos:

- **Multiplicar matrices:** en R, el operador %*% nos permite hacerlo.
- **Transponer matrices:** en R, la funcion t() nos permite hacerlo.
- **Crear matriz kxk con 0 en la diagonal principal:** en R, la función diag() nos permite hacerlo.

A tal efecto, vamos a crear nuestra matriz de residuos:

```
# Matriz de cargas
Lambda <- food.fa$loadings

# Matriz de singularidades
Psi <- diag(food.fa$uniquenesses)

# Matriz de correlaciones observada
S <- food.fa$correlation

# Creación de la matriz de correlaciones ajustada
Sigma <- Lambda %*% t(Lambda) + Psi
```

Una vez **tenemos la matriz de correlaciones observada y la matriz de correlaciones ajustada, podemos proceder a hacer una diferencia de matrices para ver el resultado que andamos buscando**

(la matriz de residuos):

```
# Creación de la matriz de residuos
round(S - Sigma, 6)
```

```
                Oil   Density    Crispy  Fracture  Hardness
Oil       0.000000  0.000001 -0.002613 -0.018220 -0.000776
Density   0.000001  0.000000 -0.001081 -0.007539 -0.000320
Crispy   -0.002613 -0.001081  0.000000  0.000000  0.000005
Fracture -0.018220 -0.007539  0.000000  0.000000  0.000033
Hardness -0.000776 -0.000320  0.000005  0.000033  0.000000
```

La matriz que ves arriba es la matriz de residuos. Su interpretación es sencilla: **cuanto más cerca estén sus valores a 0, mejor será la representación del concepto subyacente.**

3.4 INTERPRETACIÓN DE LOS FACTORES

El propósito de una rotación es producir factores con una mezcla de cargas altas y bajas y pocas cargas de tamaño moderado.

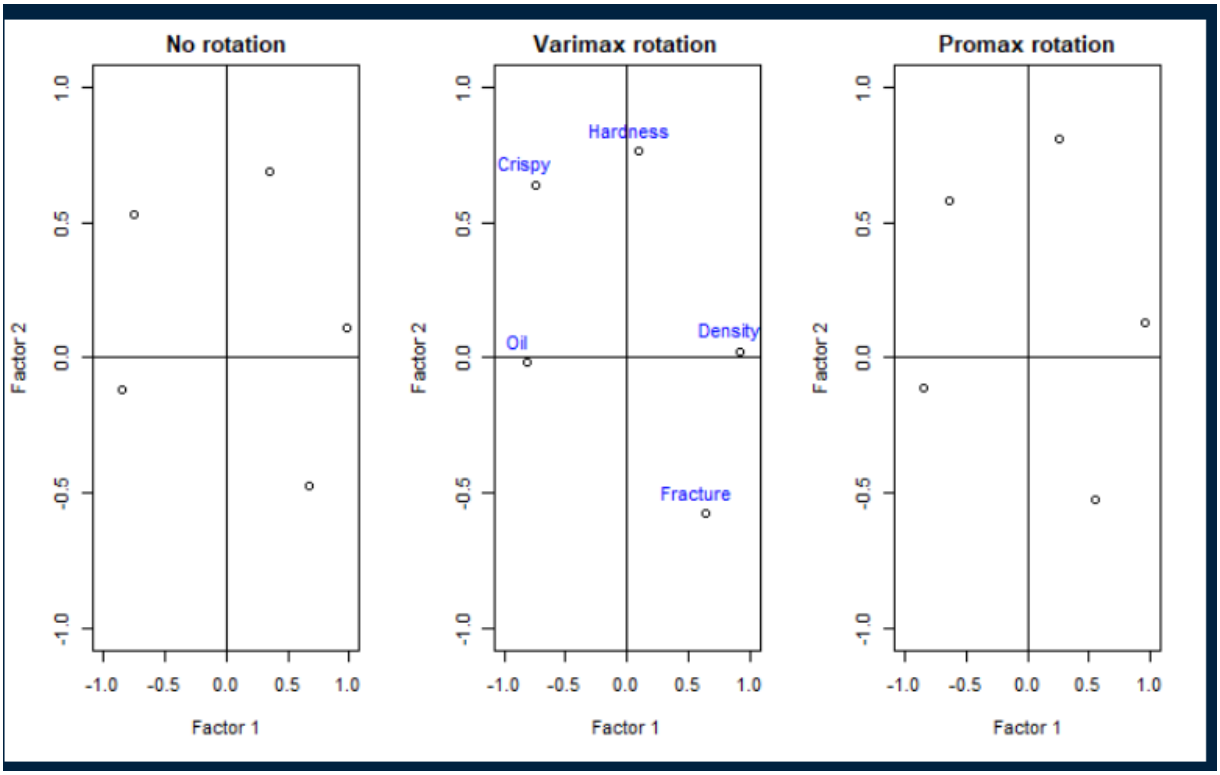
La idea es dar un significado a los factores, lo que ayuda a interpretarlos.

Desde un punto de vista matemático, no hay diferencia entre una matriz rotada y una no rotada. El modelo ajustado es el mismo, las singularidades son las mismas y la proporción de la varianza explicada es la misma.

Ajustemos tres modelos factoriales, uno sin rotación, uno con rotación varimax y uno con rotación promax, y hagamos una gráfica de dispersión de la primera y la segunda carga:

```
# Creación de 3 modelos distintos modificando la rotación
food.fa.none <- factanal(food, factors = 2, rotation = "none")
food.fa.varimax <- factanal(food, factors = 2, rotation = "varimax")
food.fa.promax <- factanal(food, factors = 2, rotation = "promax")

# Definición del output gráfico (3 gráficos en 1 fila)
par(mfrow = c(1,3))

# Primer gráfico: sin rotación
plot(food.fa.none$loadings[,1],
     food.fa.none$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "No rotation")
abline(h = 0, v = 0)

# Segundo gráfico: rotacion = varimax
plot(food.fa.varimax$loadings[,1],
     food.fa.varimax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation")
```

```r
# Texto de color azul para el gráfico segundo
text(food.fa.varimax$loadings[,1]-0.08,
     food.fa.varimax$loadings[,2]+0.08,
      colnames(food),
      col="blue")
abline(h = 0, v = 0)


# Tercer gráfico: rotacion = promax
plot(food.fa.promax$loadings[,1],
     food.fa.promax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Promax rotation")
abline(h = 0, v = 0)
```



Ahora viene la parte más divertida de todas: interpretar el significado de los factores.

Si dos variables tienen grandes cargas para el mismo factor, entonces sabemos que tienen algo en común. Observando los gráficos arriba expuestos, podremos ver que el factor 1 representa aquellos tipos de masa densas que se pueden doblar mucho (Fracture & Density), mientras que el factor 2 explica aquellos tipos de pasta que son crujientes y difíciles de romper (Crispy & Hardness).

Así que, si tuvieramos que darle un nombre a estos factores, podríamos llamarlos pastelería blanda (factor 1) y pastelería dura (factor 2).

IDEAS CLAVE

- Los Componentes Principales se construyen como combinaciones lineales de las variables iniciales y son poco interpretables.
- El algoritmo del PCA siempre intentará poner el máximo de información posible en el primer componente, luego el máximo de información restante en el segundo y así, sucesivamente.
- Estandarizar las variables es un proceso crítico para el PCA.
- La idea básica detrás del Análisis Factorial es la de una regresión, lo que significa que expresamos cada una de las variables observadas como combinaciones lineales de variables (o factores) latentes.
- La principal diferencia entre el PCA y el AF es que este último especifica, explícitamente, un modelo que relaciona las variables observadas con un conjunto más pequeño de factores subyacentes no observables mientras que el PCA no.