

Fastbook 09

Estadística Aplicada al Marketing

Reducción de dimensionalidad



09. Reducción de dimensionalidad

Con el fastbook anterior nos adentramos en los primeros modelos analíticos, los cuales irás ampliando a lo largo de todo el curso.

- Empezábamos conociendo una de las **clasificaciones de las técnicas de machine learning** basadas en la presencia o ausencia de la variable objetivo: aprendizaje supervisado frente a aprendizaje no supervisado. A su vez, estas se subdividen según el tipo de modelo a aplicar según queramos predecir o clasificar para el aprendizaje supervisado, o segmentar o reducir dimensionalidad para el aprendizaje no supervisado.
- A continuación, comprendimos el significado de **la variable objetivo o respuesta y la variable independiente o regresora**. La variable objetivo es aquella que se quiere explicar y es la que determina qué es aprendizaje supervisado; el tipo de variable, numérica o categórica, es la que determina el tipo de aprendizaje. Las variables independientes son aquellas con las que se explica la variable objetivo.
- Profundizamos en la **regresión lineal simple** (cuando solo hay una variable independiente) y **múltiple** (cuando hay más de una variable independiente). Es la técnica más utilizada cuando la variable respuesta es numérica. De esta técnica cabe destacar:
 - **Coeficientes.** Los coeficientes determinan la importancia de las variables.
 - **P-valor.** El p-valor mide la relevancia de la variable en el modelo.

- **R².** El R², o R² ajustado para regresión lineal múltiple, mide el ajuste del modelo sobre la realidad.
 - **Multicolinealidad.** Con la multicolinealidad identificamos problemas de relación entre las variables independientes.
-
- Conocemos la regresión logística cuando la **variable respuesta es binaria**. De esta técnica cabe destacar:
 - **Coeficientes.** Los coeficientes determinan la importancia de las variables.
 - **P-valor.** El p-valor mide la relevancia de la variable en el modelo.
 - **Matriz de confusión y accuracy.** Con la matriz de confusión medimos la diferencia en términos de volumen entre predicho-real y con el accuracy el porcentaje de acierto.
 - **Curva ROC y AUC.** La curva ROC es un gráfico para medir la bondad del modelo que se complementa con el AUC para tener una medida numérica de ajuste.
 - **AIC.** Métrica de ajuste para medir la calidad del modelo que pondera entre complejidad y ajuste.

Con este fastbook comprenderemos una de las dos patas del aprendizaje no supervisado: la reducción de dimensionalidad.

Dentro de reducción de dimensionalidad profundizaremos en el análisis de componentes principales, aunque también conoceremos la técnica del análisis factorial. Aprenderemos la utilidad de cada una de ellas, así como a saber interpretar los resultados obtenidos.

Autora: Patricia Martín González

¿Qué es la dimensión?

Análisis de componentes principales

Análisis factorial

Resumen

¿Qué es la dimensión?

X Edix Educación

Según la RAE, la **dimensión** son cada una de las magnitudes de un conjunto que sirven para definir un fenómeno. De forma visual: la dimensión es el número de direcciones con las que queda definido un espacio. Por ejemplo, cuando hablamos de longitud estamos en un espacio unidimensional (1 dimensión), si hablamos de área o dibujos sobre un papel tenemos 2 dimensiones (bidimensional) o si hablamos de volumen o de objetos no planos tenemos 3 dimensiones (tridimensional). En general, cuando hay varias dimensiones (más de 3) se suele decir que es un **espacio multidimensional**.

En analytics, normalmente se trabaja con espacios multidimensionales en los que la dimensión viene determinada por el número de variables con las que se crea el modelo. Por tanto, cuando tenemos una base de datos con 20 variables decimos que tiene dimensión 20 o que es multidimensional.

Como podemos imaginar, trabajar con dimensiones altas (a partir de 15-20) provoca dificultades a la hora de entender el comportamiento del modelo variable a variable, al igual que para explicar los resultados. Para ello, se aplican técnicas de **reducción de dimensionalidad** que buscan reducir el número de variables independientes transformando las originales en otras. Para algunas técnicas de reducción de dimensionalidad no puede haber dos variables linealmente dependientes (es decir, que una sea combinación de la otra) o que tengan una correlación muy alta, ya que hay cálculos que no se podrían hacer con estas variables.

Los modelos de reducción de dimensionalidad sirven para coger un (gran) número de variables de nuestro dataset, y a partir de ellas generar un número menor de variables sintéticas (sin significado real), pero con gran valor informativo. ¿Para qué sirven estas técnicas?

Transformar las variables para extraer su máxima información y luego utilizarlas en algún modelo.

Identificar y eliminar variables irrelevantes o que tienen mucha relación entre ellas.

Para visualizar el comportamiento general de las variables en una dimensión observable, 2D o 3D.

Aunque existen muchas técnicas de reducción de dimensionalidad, las dos principales son el análisis de componentes principales y el análisis factorial. Para cualquiera de las dos, después de aplicar la transformación adecuada hay que deshacer el cambio para volver a las variables originales y comprender el comportamiento en la situación inicial, que es la que realmente queremos comprender.

Análisis de componentes principales

X Edix Educación

El **análisis de componentes principales** o PCA, por sus siglas en inglés (*principal component analysis*), es la técnica más extendida dentro de la **reducción de dimensionalidad**, ya que es rápida, suele funcionar bien y es fácil de entender.

El objetivo de PCA es encontrar nuevas dimensiones, llamadas componentes principales, que reduzcan el número de variables, encontrando otras nuevas como combinación lineal de las originales que maximicen la varianza explicada y la pérdida de información sea la menor posible. Las componentes principales son perpendiculares entre ellas, por lo que la correlación es 0. Matemáticamente está basado en los autovalores y autovectores.

Las nuevas variables o componentes principales (PC) se crean a través de unos **coeficientes** α_i para cada una de las variables originales (Var). Estos coeficientes suelen venir en una matriz de rotación o loadings como resultado de la aplicación de PCA. El cálculo de estas nuevas componentes principales es:

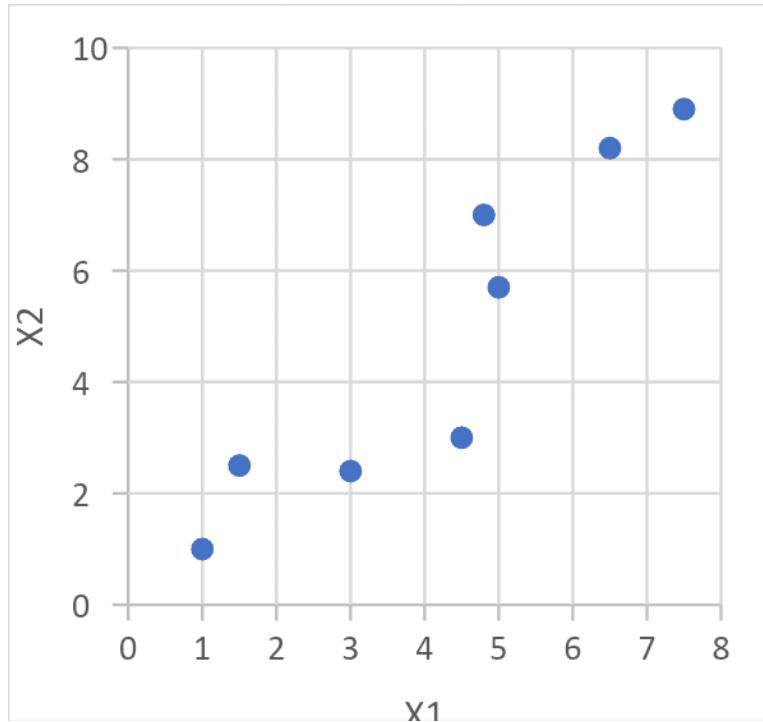
$$PC = \alpha_1 * Var_1 + \alpha_2 * Var_2 + \dots + \alpha_n * Var_n$$

Con la matriz de rotación también podemos medir la importancia de cada variable original sobre la componente principal, siendo el peso de cada variable original el valor absoluto de los coeficientes, y el impacto (positivo o negativo) el signo.

Antes de aplicar PCA a los datos, es necesario comprobar que todas las variables están en el mismo rango de valores (decenas, miles, milésimas, etc.). En caso contrario, es necesario estandarizar las variables restando la media y dividiendo por la desviación estándar.

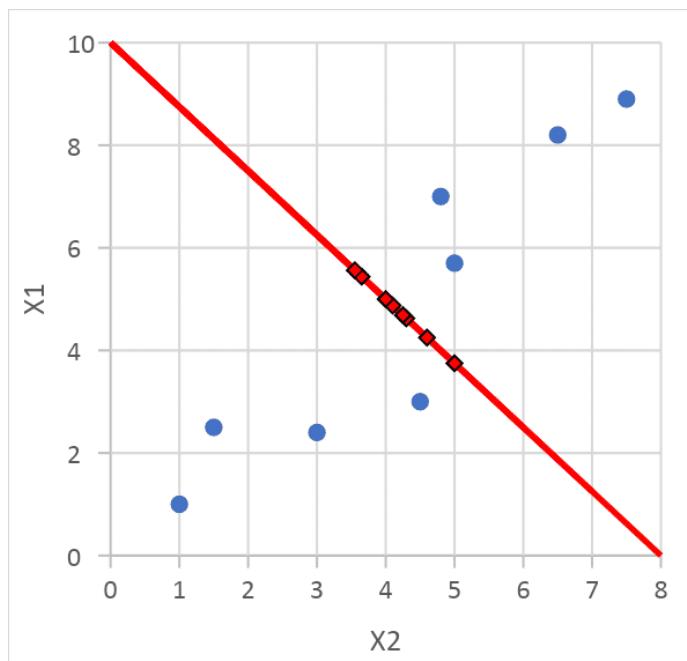
Podemos pensar visualmente en PCA como un pliegue de las variables originales en las direcciones de máxima varianza. De esta forma, los componentes principales (las nuevas variables) tienen la mayor varianza posible, lo cual suele mejorar el análisis o modelo que se haga a posteriori.

Veamos un ejemplo: supongamos que estamos en un espacio bidimensional y queremos transformarlo en unidimensional. La distribución original de nuestros puntos es:



En nuestro caso, ambas variables tienen el mismo rango y las mismas unidades por lo que no es necesario estandarizar.

Como queremos una única dirección, tenemos que ‘plegar’ las direcciones originales para buscar la dirección (componente principal) que maximice la varianza explicada. Supongamos el siguiente pliegue:



La línea roja muestra la nueva dimensión (componente principal), y los puntos rojos los nuevos puntos.

¿Crees que maximiza la varianza?

Sí

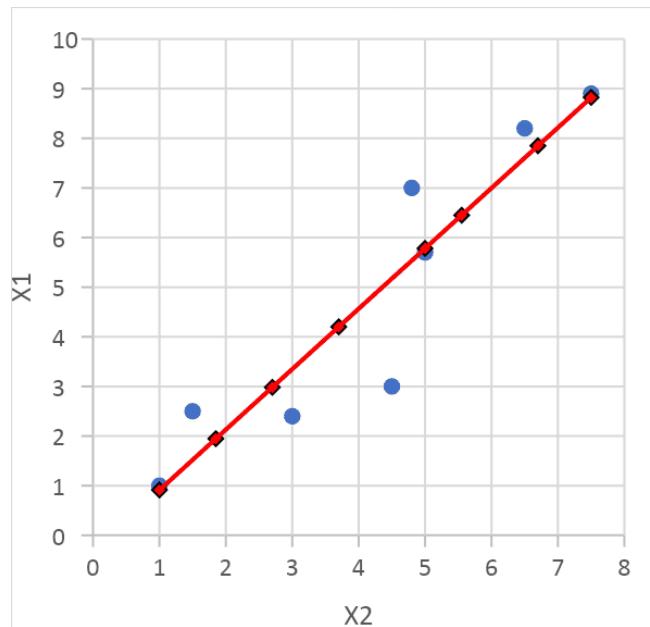
No



Respuesta: No

La componente principal no tiene demasiada varianza y no recoge correctamente el comportamiento de los puntos ya que todos están muy cerca, por lo que no es una dimensión adecuada. Eso es debido a que la dirección elegida no es la de máxima varianza.

Si aplicamos PCA, la nueva dimensión sería:



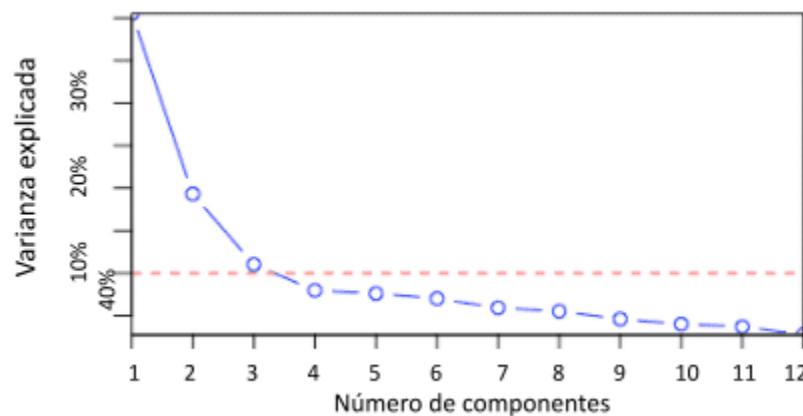
Como ves, ahora los nuevos puntos (puntos rojos) están más separados y la varianza explicada es máxima, por lo que la línea roja será la componente principal.

Como hemos visto, con el análisis de componentes principales se crean nuevas dimensiones que aportan la máxima información posible de las variables originales, pero en menos dimensiones (en este caso pasando de 2 a 1).

La varianza total explicada es el 100% y corresponde con la elección de todos los componentes principales, es decir, se seleccionan tantos componentes principales como variables originales hay. La varianza del modelo explicado por las componentes principales es la suma de las varianzas explicadas por cada una de las componentes principales. La primera componente principal será la que mayor varianza explique, la segunda componente principal, la segunda que explica mayor varianza, y así sucesivamente. Al ser todas ellas perpendiculares, no hay solapamiento y cada una explica una parte diferente.

Para la elección del número de componentes principales se suele utilizar el método del codo (con un gráfico llamado *scree plot*) para tener una estimación, aunque también es importante conocer las necesidades del problema que se esté resolviendo para determinar el número óptimo en base a la variabilidad explicada que se puede perder. En general, suele cogerse un número de componentes principales que explique al menos el 70% de la varianza total.

El diagrama del codo consiste en graficar la varianza explicada por cada una de las componentes principales, y cortar en aquel punto en el que agregar una nueva componente no sea significativo.

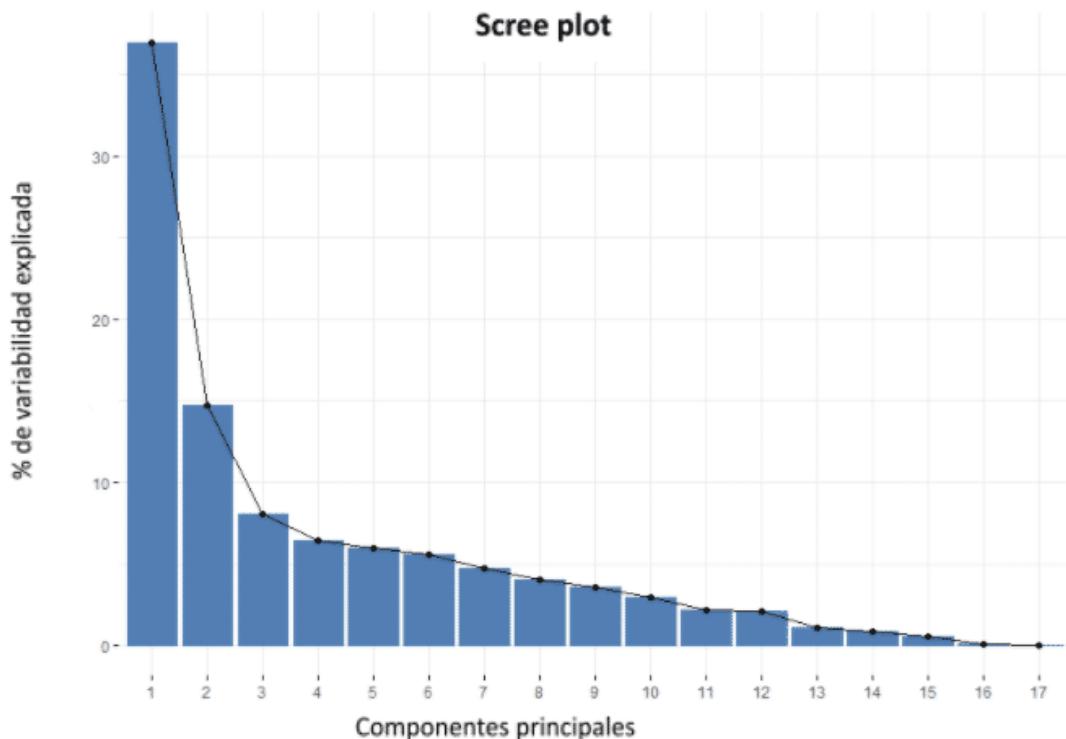


Como vemos, la línea azul se empieza a aplanar entre la tercera y la cuarta componente principal. Las tres primeras componentes principales explican el 72%, y las cuatro primeras el 80%, por lo que al ser pocas componentes es mejor quedarse con cuatro y aumentar casi un 10% la varianza explicada.

Veamos otro ejemplo. Queremos estudiar el comportamiento de los usuarios de nuestra plataforma web para saber si serán clientes muy valiosos dentro de 2 años (análisis de customer lifetime value). Seleccionamos 17 variables incorreladas, pero siguen siendo muchas, por lo que aplicamos PCA.

Entre las salidas de las componentes principales, tenemos la varianza explicada por cada una de ellas y la matriz de rotación o loadings para crear las componentes principales.

El porcentaje de varianza explicada de cada una de las componentes principales es:



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Varianza explicada	37,0%	14,8%	8,1%	6,5%	6,0%	5,6%	4,8%	4,0%	3,6%
Varianza explicada acumulada	37,0%	51,7%	59,8%	66,3%	72,3%	77,9%	82,6%	86,7%	90,2%

	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Varianza explicada	2,9%	2,1%	2,1%	1,1%	0,9%	0,6%	0,1%	0,0%
Varianza explicada acumulada	93,2%	95,3%	97,4%	98,5%	99,3%	99,9%	100,0%	100,0%

Como vemos en el gráfico, a partir de la 4 componente principal se estabiliza la varianza explicada, por lo que podríamos elegir 3 como número de componentes principales adecuados. Si completamos la información con la tabla inferior, observamos que la varianza explicada acumulada llega al 70% con la componente principal 5 (PC5), por lo que seleccionamos 5 como número óptimo de componentes principales.

Las **nuevas dimensiones**, como hemos dicho anteriormente, se crean como combinación lineal de las variables originales. La matriz de rotación o loadings contiene los coeficientes para calcular las nuevas dimensiones:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	...
Var1	-0,38	0,03	-0,05	0,15	-0,04	0,01	-0,04	...
Var2	-0,30	0,25	0,06	-0,17	0,06	-0,02	-0,03	...
Var3	-0,10	-0,16	0,56	-0,06	0,19	0,29	0,07	...
Var4	0,01	-0,05	0,10	-0,07	-0,85	0,47	-0,15	...
Var5	-0,26	0,37	0,01	0,03	-0,06	0,05	0,11	...
Var6	-0,32	-0,27	-0,11	0,06	-0,05	0,01	-0,09	...
Var7	-0,30	0,20	0,09	-0,16	0,07	-0,04	-0,06	...
Var8	-0,14	-0,22	-0,06	-0,04	-0,18	-0,09	0,92	...
Var9	-0,25	-0,31	0,03	0,21	-0,07	-0,15	-0,04	...
Var10	-0,07	-0,18	0,56	-0,17	0,25	0,25	0,03	...
Var11	-0,36	0,21	-0,02	0,07	-0,02	0,04	0,01	...
Var12	-0,26	0,44	0,00	-0,13	-0,02	0,07	0,10	...
Var13	-0,23	-0,32	-0,28	-0,25	0,08	0,06	-0,12	...
Var14	-0,16	-0,26	-0,35	-0,38	0,16	0,36	-0,03	...
Var15	-0,11	-0,15	0,34	-0,30	-0,31	-0,65	-0,13	...
Var16	-0,21	-0,13	0,09	0,72	0,03	0,09	-0,04	...
Var17	-0,27	-0,19	-0,08	0,00	0,00	-0,18	-0,23	...

En las filas tenemos las variables originales y en las columnas los componentes principales. Por tanto, los componentes principales se crean como:

$$PC1 = -0,39 * Var1 - 0,30 * Var2 - 0,10 * Var3 + 0,01 * Var4 + \dots - 0,27 * Var17$$

$$PC2 = 0,03 * Var1 + 0,25 * Var2 - 0,16 * Var3 - 0,05 * Var4 + \dots - 0,19 * Var17$$

En nuestro caso solamente crearíamos los 5 primeros componentes principales como hemos visto anteriormente.

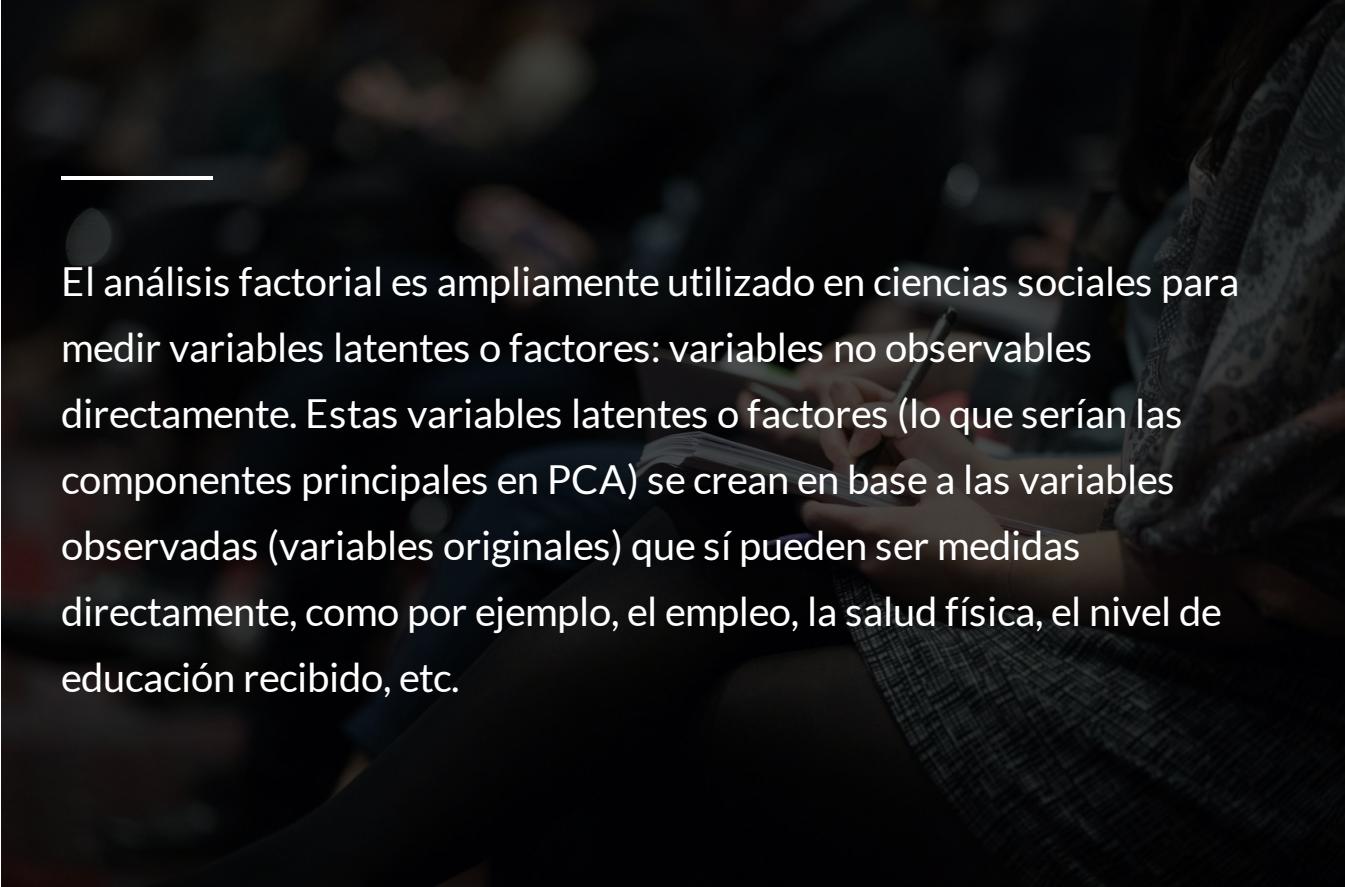
Para medir la importancia de cada variable original sobre las componentes principales miramos el valor absoluto de los coeficientes, el tipo de impacto en el signo.

En PC1 las variables Var1 y Var11 son las que mayor peso tienen con impacto negativo, mientras que Var4 tiene prácticamente impacto 0. Para PC2, la variable con mayor impacto es la Var5 con impacto positivo, mientras que las Var1 y Var4 apenas tienen impacto.

Análisis factorial

X Edix Educación

El análisis factorial o FA por sus siglas en inglés (*factor analysis*) tiene la misma estructura que el análisis de componentes principales, aunque con alguna diferencia: PCA es un método descriptivo para reducir la dimensionalidad, mientras que el análisis factorial está pensado como técnica para el análisis de intangibles (intangible es todo aquello que no puede ser percibido físicamente).

A dark, slightly blurred photograph showing a person's hands holding a pen and writing in a notebook. The notebook has a grid pattern on its cover. The hands are positioned as if in the middle of writing a sentence.

El análisis factorial es ampliamente utilizado en ciencias sociales para medir variables latentes o factores: variables no observables directamente. Estas variables latentes o factores (lo que serían las componentes principales en PCA) se crean en base a las variables observadas (variables originales) que sí pueden ser medidas directamente, como por ejemplo, el empleo, la salud física, el nivel de educación recibido, etc.

El análisis factorial crea factores basados en la correlación de las variables originales. Hay varias formas de calcularlos: maximizar la varianza, minimizar el número de factores, crear variables latentes oblicuas en lugar de ortogonales, Varimax (crea factores que no tengan saturaciones altas de las variables originales)... Se elegirá un método en función de las necesidades, aunque el más extendido es Varimax.

A diferencia de PCA, donde se medía la explicabilidad total de la BBDD original, **en FA se mide el porcentaje de explicabilidad de las variables originales** en una variable llamada comunalidad como la suma de los cuadrados de los factores:

$$\text{comunalidad} = \text{Factor1}^2 + \text{Factor2}^2 + \dots + \text{Factorn}^2$$

Veamos un **ejemplo**. Consideremos los datos socioeconómicos de los países para entender el comportamiento oculto de los datos.

	Tasa de natalidad	Tasa de mortalidad	Mortalidad infantil	Esperanza de vida (H)	Esperanza de vida (M)	PIB
País 1	24,7	5,7	30,8	69,6	75,5	600
País 2	12,5	11,9	14,4	68,3	74,7	2.250
País 3	13,4	11,7	11,3	71,8	77,7	2.980
País 4	11,6	13,4	14,8	65,4	73,8	2.780
País 5	14,3	10,2	16,0	67,2	75,7	1.690
País 6	13,6	10,7	26,9	66,5	72,4	1.640
País 7	17,7	10,0	23,0	64,6	74,0	2.242
País 8	15,2	9,5	13,1	66,4	75,9	1.880
País 9	13,4	11,6	13,0	66,4	74,8	1.320

Al GNP (PIB) le aplicamos el logaritmo para conseguir una distribución más parecida al resto de variables, y estandarizamos. Aplicamos FA, cogemos dos factores, y obtenemos la matriz loadings:

	Factor 1	Factor 2
Tasa de natalidad	-0,8900	0,2217
Tasa de mortalidad	-0,3340	0,9399
Mortalidad infantil	-0,8608	0,4154
Esperanza de vida (H)	0,8505	-0,5000
Esperanza de vida (M)	0,8939	-0,4428
Log (PIB)	0,7814	-0,2810

Como vemos, en el primer factor influyen todas las variables excepto la tasa de mortalidad, por lo que podemos decir que la variable latente 1 mide el desarrollo del país. En el segundo factor, el mayor peso lo tiene la tasa de mortalidad, seguido de la esperanza de vida de los hombres. Podemos decir que la segunda variable mide aspectos relacionados con la salud.

Calculamos ahora la communalidad (el porcentaje de variabilidad explicada para cada variable original en función de los factores):

	Factor 1	Factor 2	Comunalidad
Tasa de natalidad	-0,8900	0,2217	84,12%
Tasa de mortalidad	-0,3340	0,9399	99,50%
Mortalidad infantil	-0,8608	0,4154	91,34%
Esperanza de vida (H)	0,8505	-0,5000	97,34%
Esperanza de vida (M)	0,8939	-0,4428	99,51%
Log (PIB)	0,7814	-0,2810	68,95%

Por ejemplo, para la tasa de nacimiento:

$$\text{comunalidad}_{\text{tasa de natalidad}} = -0,8900^2 + 0,2217^2 = 0,8412$$

A la vista de los resultados, vemos que casi todas las variables quedan muy explicadas excepto el logaritmo del PIB del que conseguimos explicar un 69%.

Resumen

X Edix Educación

En este fastbook, hemos comprendido el significado de dimensión y la importancia que tiene para interpretar y comprender los resultados obtenidos cuando es muy alta. Para reducir este problema, hemos conocido las dos principales técnicas de reducción de dimensionalidad: el análisis de componentes principales (PCA) y el análisis factorial (FA).

El análisis de componentes principales o PCA

Es una técnica descriptiva que crea las nuevas variables maximizando la **varianza explicada**. Estas nuevas variables se crean como combinación lineal de las variables originales multiplicando los coeficientes que se obtienen en la matriz de loadings o de rotación. El número de componentes principales se eligen apoyados en el **diagrama o gráfico del codo** y las necesidades de negocio. En general, se suele escoger un número de componentes principales que explique, al menos, el **70%** de la varianza.

El análisis factorial o FA

Es una técnica utilizada sobre todo en el ámbito social para crear variables que no son observables directamente. Las nuevas variables, llamadas **variables latentes** o **factores**, se crean como combinación lineal de las variables originales basadas en la correlación entre ellas, aunque su método de cálculo puede ser muy diverso. A diferencia de PCA, en el análisis factorial se mide la varianza explicada de las variables **originales** como suma de cuadrados de los **factores o variables latentes**.

¡Enhорabuena! Fastbook superado

edix

Creamos Digital Workers