
Componentes Principales

Pedro Valero Mora-valerop@uv.es

Metodología de las CC del Comp-Universitat de València

Marzo 2012



UNIVERSITAT DE VALÈNCIA

Contenidos

Introducción

El problema

Métodos de reducción de la dimensionalidad

Objetivo

Pasos

1

2

4

5

6

Ejemplo: Crímenes

Crímenes

Input

Output

El número de componentes

7

8

9

14

21

Introducción

El problema

- Cuando uno tiene muchas variables muy redundantes, hacer análisis una por una es excesivo
 - Es conveniente combinar esas variables de modo que tengamos un resumen
- Una forma de combinar es simplemente hacer la suma de las preguntas
 - Un cuestionario con 12 variables sobre satisfacción corporal podría calcularse la suma de las preguntas para obtener una puntuación de satisfacción corporal
 - Sin embargo, esa suma puede ser un sinsentido si las preguntas no correlacionan entre sí
 - Es habitual que un conjunto de preguntas en un cuestionario no todas ellas correlacionen entre sí y que se pueda distinguir entre dos o tres conjuntos de variables que correlacionan
 - Esos conjuntos formarían subescalas que se podrían combinar

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

Crímenes
Input
Output
El número de componentes
[Actividades](#)

- Otra aplicación es la visualización de muchas variables
 - Gráficos de los componentes principales de una matriz con muchas variables permiten ver buena parte de la varianza original, lo cual es imposible con diagramas de dispersión u otro tipo de gráficos

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

Crímenes
Input
Output
El número de componentes
[Actividades](#)

Métodos de reducción de la dimensionalidad

- Existen dos métodos básicos de reducción de la dimensionalidad
 - Análisis de componentes principales
 - Análisis factorial
- Ambos son muy parecidos y tienden a confundirse
 - Aquí nos centraremos en análisis de componentes principales por razones de tiempo y porque es una buena introducción al análisis factorial

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

Crímenes
Input
Output
El número de componentes
[Actividades](#)

Objetivo

- El ACP permite
 - Representar las relaciones en un conjunto de variables de una manera más simple
 - Simplificar el conjunto de datos, combinando variables muy correlacionadas y produciendo componentes independientes entre sí
 - Cuando los factores son interpretables, estos pueden ayudarnos a entender los datos de una manera diferente
 - Combinar variables que tengan relación (frente a simplemente a sumar)
- En ACP no hace falta especificar un número de factores y a menudo se utiliza como un método exploratorio previo a hacer análisis factorial

Pasos

- Los pasos suelen ser:
 - Calcular una matriz de correlaciones entre las variables
 - Extraer los factores y determinar el número. También diagnosticar el ajuste
 - Rotación de factores (no lo veremos)
 - Calculo de puntuaciones que sean una combinación de las variables más relacionadas

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

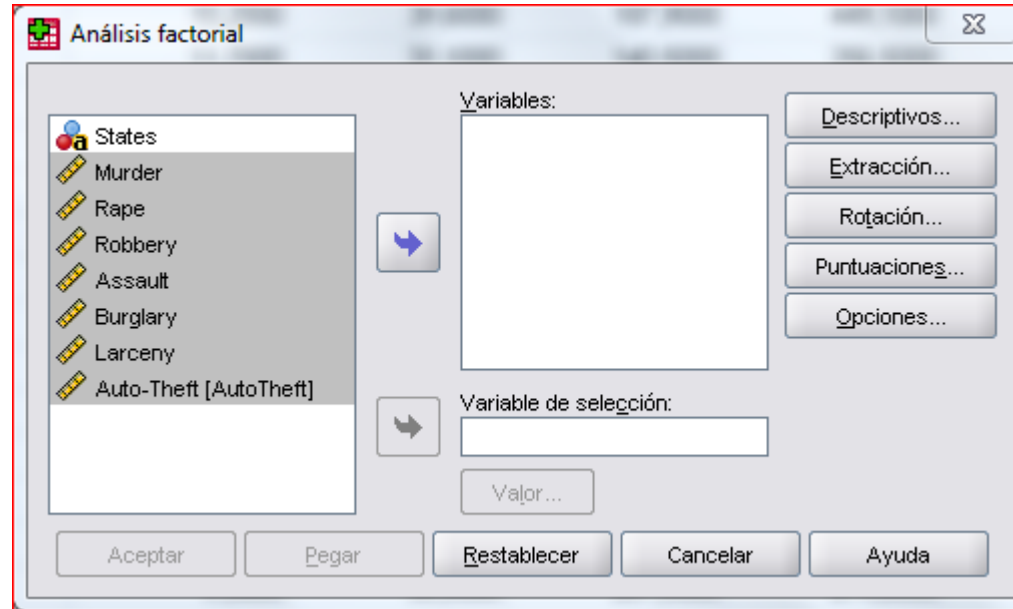
Crímenes
Input
Output
El número de componentes
[Actividades](#)

Ejemplo: Crímenes

Crímenes

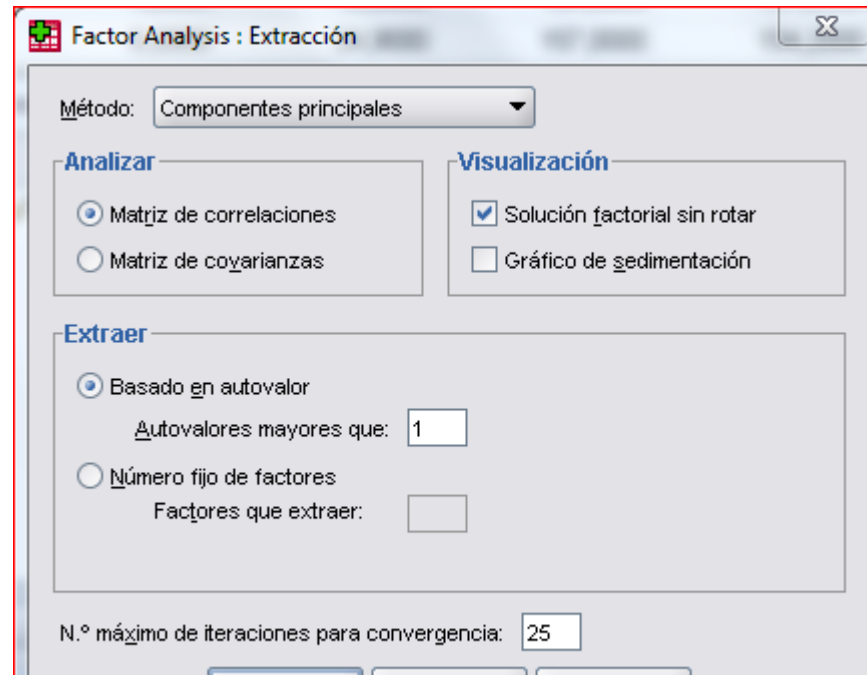
- Se trata del número de crímenes por 100.000 habitantes en los años 80 en USA por estado en una serie de categorías de crímenes
- La idea es reducir los tipos de crímenes a un número más reducido de variables
- También, en este caso es interesante examinar las posiciones de los estados individuales y ver qué factores los caracterizan

Input



- El primer paso es elegir las variables

- En Extracción hay opciones importantes

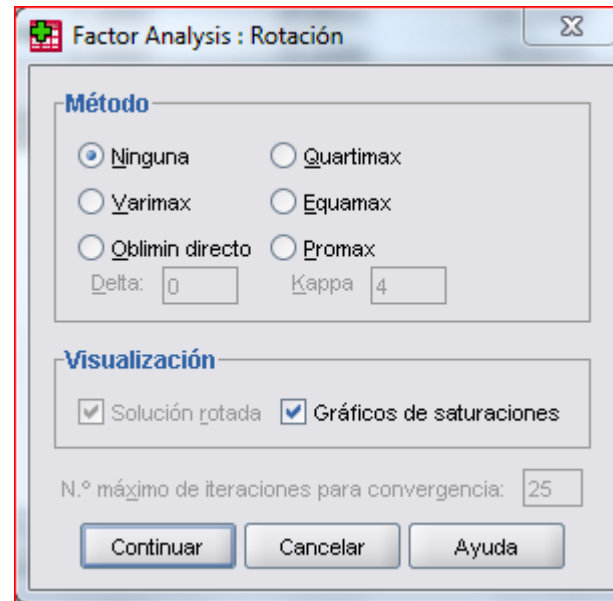


- Método: Usaremos componentes principales
- Analizar: Usaremos correlaciones (equivale a estandarizar). Covarianzas solo tiene sentido cuando la varianza es similar (por ejemplo, ítems cuestionario)
- Visualización: Es conveniente elegir el gráfico de sedimentación
- Extraer: Autovalores mayores que 1 para mí es poco (yo lo bajaría pero vamos a ver los valores por defecto)

El problema
Métodos de reducción de la
Objetivo
Pasos

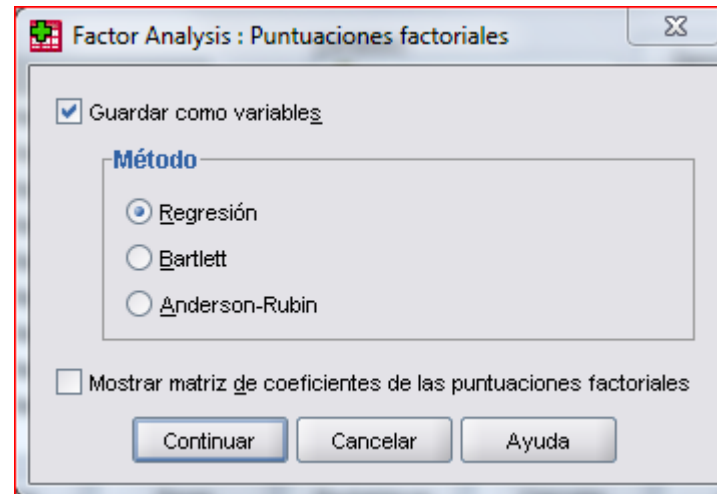
Crímenes
Input
Output
El número de componentes
[Actividades](#)

- Rotación y visualización



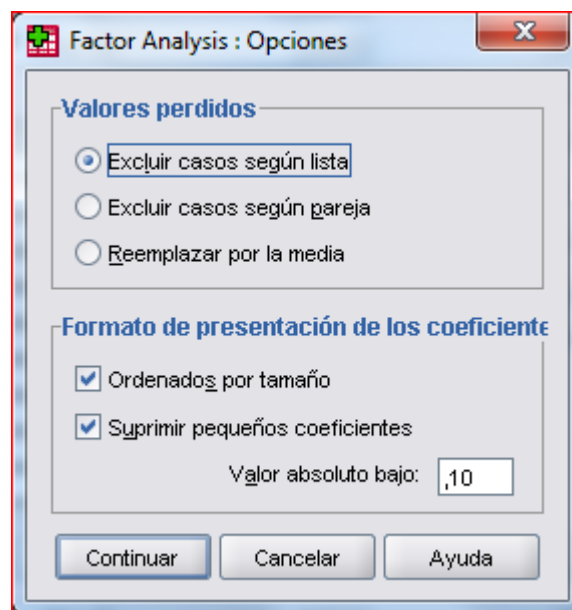
- Sobre la rotación, no las veremos aquí.
- En este caso es interesante el gráfico de saturaciones si el resultado tiene sólo dos componentes (con más de dos el gráfico es muy flojo)

- Puntuaciones factoriales



- Si se elige componentes principales en el método, da igual el método para calcular las puntuaciones factoriales (todas dan lo mismo)
- La matriz de coeficientes sirve para ver la fórmula que permite calcular las puntuaciones a partir de las puntuaciones originales en las variables (podría usarse para obtener una puntuación que no ha sido usada en el análisis)

- Opciones



- El formato de presentación de los coeficientes puede ser útil para explorar los resultados
 - Ordenados por tamaño: Realmente sólo afecta a la primera columna
 - Suprimir pequeños coeficientes: Limpia la matriz de coeficientes y quita los valores pequeños

Output

- Comunalidades

Communalities		
	Initial	Extraction
Murder	1,000	,861
Rape	1,000	,803
Robbery	1,000	,650
Assault	1,000	,794
Burglary	1,000	,848
Larceny	1,000	,726
Auto-Theft	1,000	,671

Extraction Method: Principal

Component Analysis.

- Como usamos componentes principales, la inicial es 1
- En esta parte vemos el tanto por ciento de varianza que ha sido explicado con el número de componentes considerado (2, veremos porqué más adelante)
- Si en el input hubieramos pedido que considerara más componentes la extracción podría llegar a 1 que es el máximo
- Nos da una idea de cómo es de buena la solución para las variables consideradas

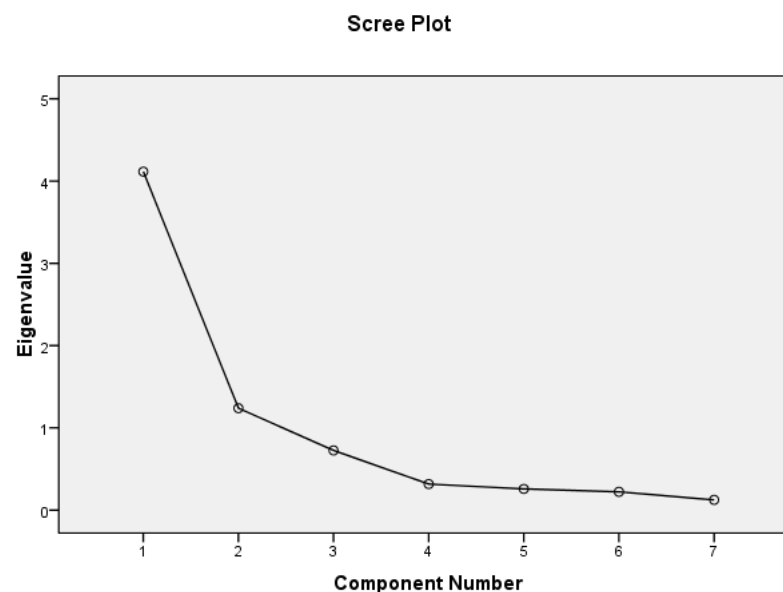
- Varianza total explicada

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,115	58,785	58,785	4,115	58,785	58,785
2	1,239	17,696	76,481	1,239	17,696	76,481
3	,726	10,369	86,850			
4	,316	4,520	91,370			
5	,258	3,685	95,056			
6	,222	3,172	98,228			
7	,124	1,772	100,000			

Extraction Method: Principal Component Analysis.

- Esta tabla muestra la descomposición de varianza en los diversos componentes
La suma de la columna Total es igual al número de variables
Si todas las variables fueran independientes entre sí, la columna de eigenvalores tendría unos y el ACP no tendría interés
Los primeros componentes suelen tener eigenv. altos y los últimos bajos
La siguiente columna son los porcentajes con respecto al total de la izquierda
- Puesto que sólo hemos considerado dos componentes, el resultado muestra la parte correspondiente a dos componentes en las 3 últimas columnas (el 76% de la varianza es explicada con dos componentes)

- Scree plot (gráfico de sedimentación)



- Decidir cuál es el número correcto de componentes a considerar es un problema a menudo
- El Scree plot es una solución muy antigua.

La idea es ver donde ese gráfico se dobla y empieza a tener aspecto horizontal

Sin embargo es bastante normal que ese pliegue no esté muy claro

- Matriz de componentes

Component Matrix ^a		
	Component	
	1	2
Burglary	,893	,226
Rape	,876	-,189
Robbery	,805	
Assault	,805	-,382
Larceny	,725	,448
Auto-Theft	,599	,559
Murder	,609	-,700

Extraction Method: Principal

Component Analysis.

a. 2 components extracted.

- Esta matriz se puede interpretar como la correlación de cada variable con cada componente

Analizando estas correlaciones podemos dar significado a los componentes obtenidos

- Forzando el análisis a mostrar todos los componentes podemos entender mejor esta tabla:

	Component						
	1	2	3	4	5	6	7
Murder	.609	-.700	.152	-.131	.273	.122	.094
Rape	.876	-.189	-.208	.035	.096	-.364	-.104
Robbery	.805	.047	.422	-.314	-.264	-.054	-.001
Assault	.805	-.382	-.059	.354	-.257	.081	.068
Burglary	.893	.226	-.179	-.032	.051	.253	-.228
Larceny	.725	.448	-.459	-.132	.015	.019	.212
Auto-Theft	.599	.559	.484	.236	.188	-.027	.052
	4.11495951	1.23872183	0.72581663	0.31643205	0.25797446	0.22203947	0.12405606

1
1
1
1
1
1
1

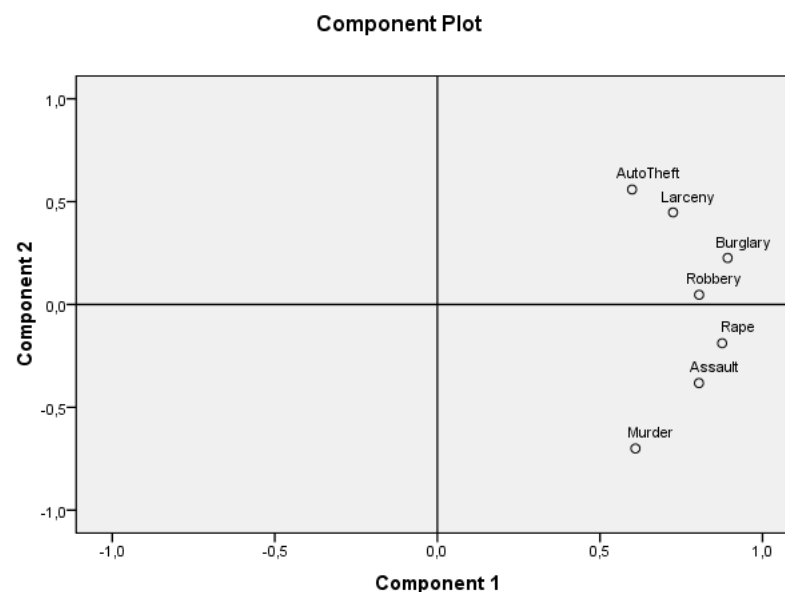
La suma de cuadrados por columnas es igual al eigenvector, y la suma por filas es igual a 1

Cada columna es la correlación del componente con la variable

La primera columna es la que tiene correlaciones más altas y las demás progresivamente tienen correlaciones más bajas

Es habitual que el primer componente sea un componente de tamaño (en este caso de mayor o menor criminalidad) y que el segundo o el tercero sea más interesante de interpretar

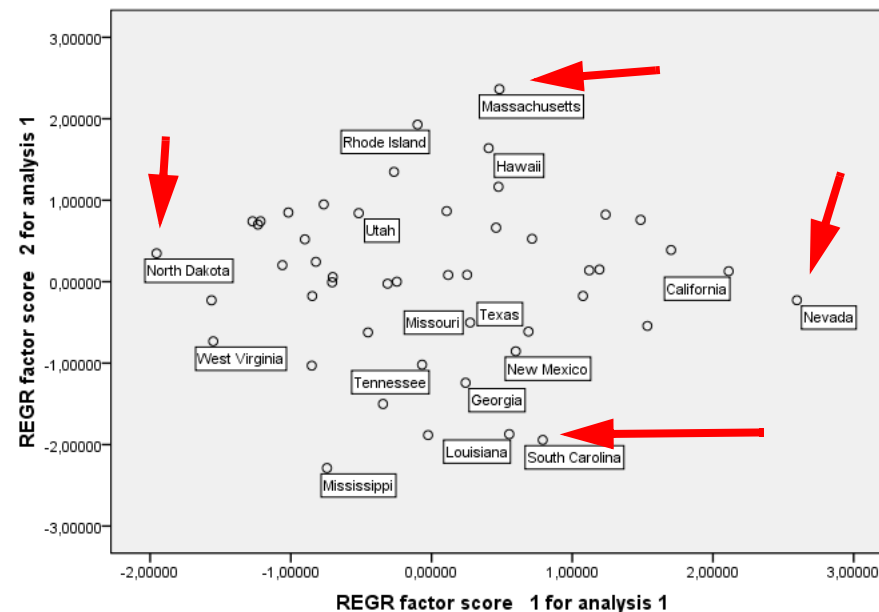
- Gráfico de componentes



- Este gráfico con dos componentes pone matriz anterior en un plano
El componente 1 está en el plano horizontal y vemos que todas las variables están en el lado positivo
El componente 2 está en el plano vertical y vemos que Autotheft, Larceny y Burglary están en el positivo, Robbery casi en el cero, y Rape, Assault y Murder en el negativo

- Gráfico de puntuaciones en los componentes
 - Si hemos elegido guardar como variables anteriormente SPSS nos añade columnas a los datos originales

Un gráfico de dispersión de las puntuaciones en los componentes puede ser interesante para interpretar (sobre todo si las puntuaciones tienen etiquetas)

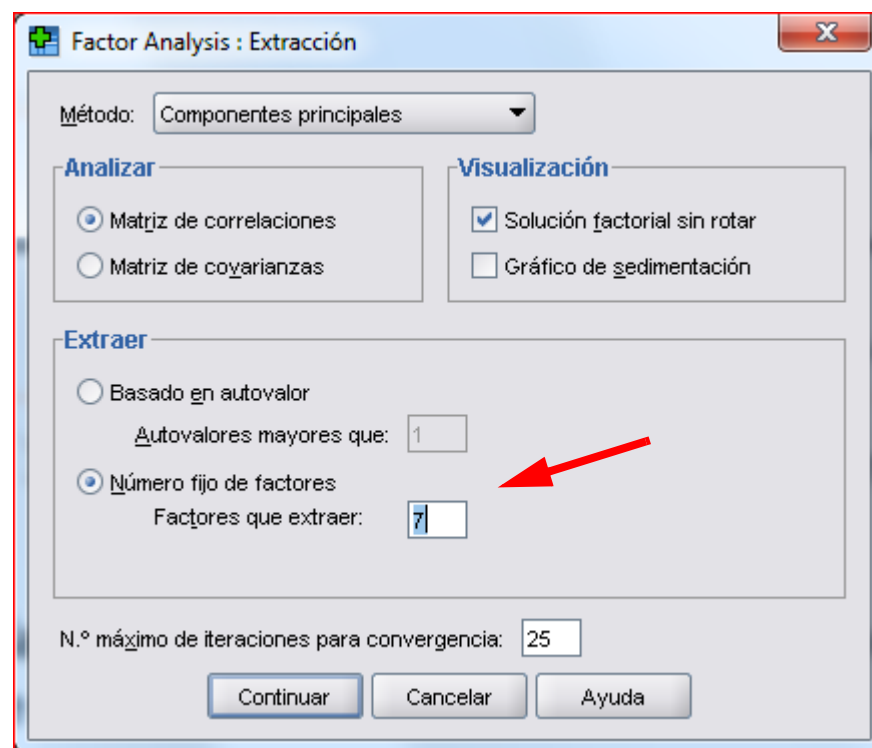


Aquí vemos que Nevada tiene muchos crímenes y Dakota del Norte no
Massach. destaca en crímenes contra la prop y S. Carolina contra las personas

El número de componentes

- Antes hemos visto que SPSS ofrece algunos métodos para reducir el número de componentes
 - En realidad, el ACP produce tantos componentes como variables, sólo que cada uno de ellos tiene cada vez menos varianza
- ¿Cuántos hay que interpretar?
 - 2 es un número interesante porque podemos ponerlos en un gráfico
 - Los que tienen eigenvalores por encima de 1 es otra posibilidad
 - Mirando la matriz de componentes, a partir de cierto momento cada componente está asociado con una sola variable. Eso representa el componente único de esa variable
 - Usar el scree plot puede ayudar a tomar la decisión
 - El análisis paralelo (el SPSS no lo hace) es interesante (Ledesma et. al 2007 para un software para hacerlo)

- Mi recomendación es que al menos 3, ya que el primero no suele ser muy interesante
 - Examinarlos todos tampoco es mala idea. Aunque los últimos tengan poca información, esa información puede ser la más interesante
- Para obtener más componentes



Actividades

1. **Sexo, Drogas y Rock&Roll.** En los datos de Musica.sav hay unos datos sobre las preferencias musicales de un grupo de encuestados. Calcula el ACP y correlaciona los componentes con las variables relaciones sexuales y ser favorable a la legalización de la marihuana.

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

Crímenes
Input
Output
El número de componentes
[Actividades](#)

2. Para los datos sobre satisfacción corporal (archivo BodySatisfactionEndurance) calcular el análisis de componentes principales para la escala de satisfacción corporal ¿Dirías que un sólo componente es suficiente para resumir esa escala? Calcula luego la correlación del primer componente después de guardar las puntuaciones con la variable Body Mass Index y con el Género del estudiante. Podeis comparar los resultados también con la puntuación de satisfacción total que está en los datos y que no es más que la suma de los ítems de la escala de satisfacción.

Introducción

El problema
Métodos de reducción de la
Objetivo
Pasos

Ejemplo: Crímenes

Crímenes
Input
Output
El número de componentes
[Actividades](#)

3. Hacer lo mismo con la escala de aguante (self endurance), ¿parece una escala de un sólo componente? Calcula las correlaciones entre los componentes de satisfacción corporal y los obtenidos para aguante (self endurance).

Introducción

- El problema
- Métodos de reducción de la
- Objetivo
- Pasos

Ejemplo: Crímenes

- Crímenes
 - Input
 - Output
- El número de componentes
- [Actividades](#)

