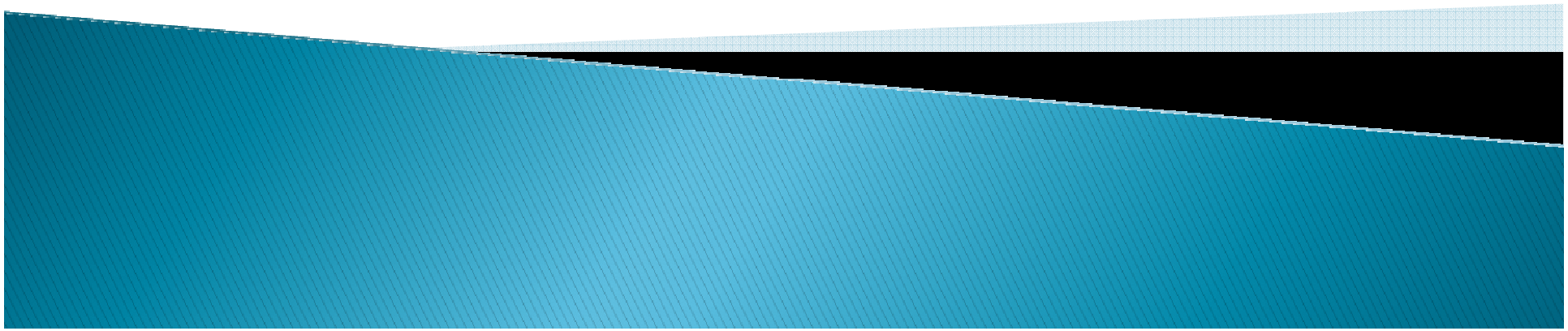
Practica 2: Componentes Principales

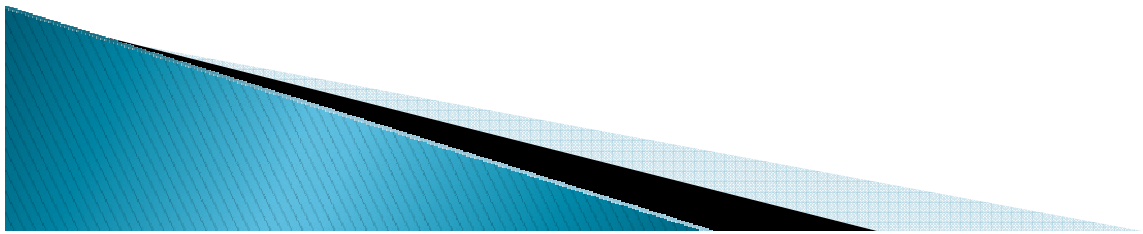
13- octubre -2009



Componentes Principales:

Nos permite transformar las variables originales (X), en general correladas, en nuevas variables incorreladas (Z), facilitando su interpretación.

De modo ideal, se buscan m $<$ p variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.



Abrir los datos en SPSS.. (paso 1 al 3)

EJEMPLO ⇒ AUTOMOVILES DE TURISMO (texto)

Datos originales: X =

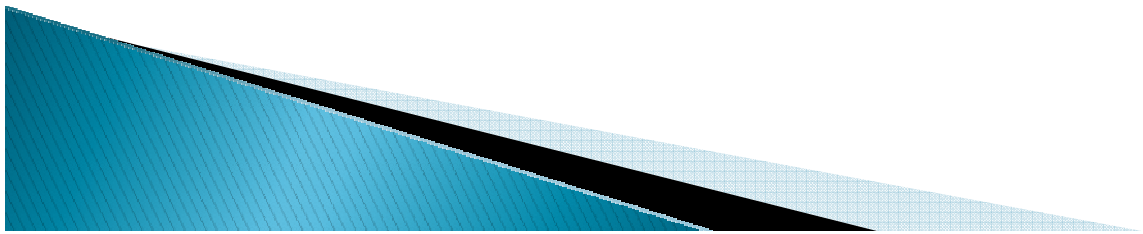
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	4	1	4	3	3	2	4	4	4	4
2	5	5	4	4	3	3	4	1	1	3
3	2	1	3	1	4	2	1	5	4	5
4	1	1	1	1	4	4	2	5	5	4
5	1	1	2	1	5	5	4	3	3	2
6	5	5	5	5	3	3	4	2	2	1
7	4	5	4	4	2	2	5	1	1	1
8	3	2	3	1	4	4	2	5	5	5
9	4	4	4	3	4	4	3	1	1	1
10	5	5	5	5	2	2	3	2	2	2
11	2	2	2	1	5	4	4	3	4	3
12	4	4	5	5	4	5	5	2	1	2
13	3	2	2	1	4	5	4	4	3	3
14	5	5	4	4	5	4	4	1	2	2
15	4	3	3	1	4	4	5	3	4	4
16	5	5	4	4	4	5	4	2	1	1
17	4	4	5	2	4	5	5	4	4	2
18	5	5	4	4	2	2	1	2	2	3
19	3	3	2	2	4	4	5	4	5	4
20	5	5	4	4	4	5	4	3	2	1

√ Los datos originales tienen escala “ordinal”, los números del 1 al 5 representan una graduación de “menos interés” a “mayor interés” (las trataremos como numéricas)

- Ejecutar todas las salidas en SPSS

Menú Inicio->Programas->SPSS Inc->PASW Statistics 18

Analizar -> Reducción de dimensionalidad -> Factor



Estadísticos descriptivos:

	Media	Desviación típica	análisis
Val.Preci	**3.70**	1.342	20
Val.Fina	3.40	1.635	20
Val.Cons	3.50	1.192	20
Val.Comb	2.80	1.576	20
Val.Segu	**3.70**	.923	20
Val.Conf	**3.70**	1.174	20
Val.Capa	**3.65**	1.268	20
Val.Prest	2.85	1.387	20
Val.Mode	2.80	1.473	20
Val.Aero	2.65	1.348	20

Para los encuestados, el precio (V1), la seguridad (V5), el confort (V6) y la capacidad (V7) son atributos más importantes que el resto.

Matriz de correlaciones

	Val.Preci	Val.Fina	Val.Cons	Val.Comb	Val.Segu	Val.Conf	Val.Capa	Val.Prest	Val.Mode	Val.Aero
Val.Preci	1.000	.873	.823	.816	-.501	-.194	.213	-.648	-.645	-.497
Val.Fina	**.873**	1.000	.729	.829	-.439	-.071	.249	-.784	-.752	-.697
Val.Cons	**.823**	**.729**	1.000	.812	-.478	-.226	.192	-.557	-.630	-.540
Val.Comb	**.816**	**.829**	**.812**	1.000	-.550	-.262	.174	-.737	-.789	-.654
Val.Segu	-.501	-.439	-.478	-.550	1.000	.738	.175	.292	.341	.123
Val.Conf	-.194	-.071	-.226	-.262	**.738**	1.000	.421	.132	.055	-.236
Val.Capa	.213	.249	.192	.174	.175	.421	1.000	-.301	-.180	-.414
Val.Prest	**-.648**	**-.784**	-.557	-.737	.292	.132	-.301	1.000	.886	.730
Val.Mode	**-.645**	**-.752**	-.630	**-.789**	.341	.055	-.180	**.886**	1.000	.785
Val.Aero	**-.497**	-.697	-.540	-.654	.123	-.236	-.414	**.730**	**.785**	1.000

El precio (V1), la financiación (V2), el consumo (V3) y el tipo de combustible (V4) están bastante correladas.

También las prestaciones (V8), la modernidad (V9) y la aerodinámica (V10).

- Tiene sentido Comp. Principales

(correlaciones en valor absoluto)

Prueba de Bartlett

El test nos sirve para comprobar que las correlaciones entre las variables son distintas de cero de modo significativo.

En general, el determinante de la matriz nos da una idea de la correlación generalizada entre todas las variables.

Var. Correladas ⇒ determinante es menor que uno
Var. Incorreladas ⇒ determinante es uno

$$H_0: |R| = 1$$
$$H_1: |R| \neq 1$$

Quisiéramos rechazar..

En SPSS, tenemos:

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		.700
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	163.466
	gl	45
	Sig.	**.000**

Se rechaza la hipótesis de que las variables son incorreladas, es decir, las variables son correladas significativamente.

Teorema 3.1 Sea S_Y la matriz de covarianzas asociada al vector $\mathbf{Y} = (Y_1, \ldots, Y_p)'$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ los valores propios de S_Y, con vectores propios asociados $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_p$. Se verifica:

(i) La componente principal i-ésima es

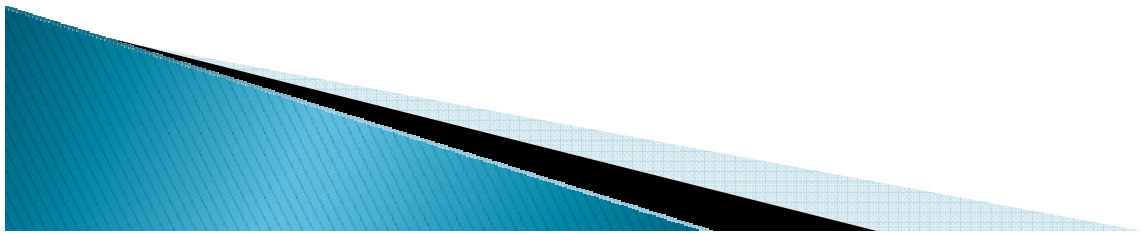
$$Z_i = \mathbf{a}_i'\mathbf{Y} = a_{1i}Y_1 + a_{2i}Y_2 + \cdots + a_{pi}Y_p, \;\; i = 1, \ldots, p\,.$$

(ii) La varianza de la componente principal i-ésima es

$$s^2_{Z_i} = \;\lambda_i, \;\; i = 1, \ldots, p\,.$$

(iii) La covarianza entre dos componentes Z_i y Z_j es

$$s_{Z_iZ_j} = 0, \;\; \forall i \neq j\,.$$



Comunalidades

Comunalidades				
	Bruta		Reescalada	
	1 Inicial	2 Extracción	Inicial	3 Extracción
Val.Preci	1,800	1,414	1,000	,786
Val.Fina	2,674	2,328	1,000	,871
Val.Cons	1,421	1,002	1,000	,705
Val.Comb	2,484	2,218	1,000	,893
Val.Segu	,853	,606	1,000	,711
Val.Conf	1,379	1,103	1,000	,800
Val.Capa	1,608	,926	1,000	,576
Val.Prest	1,924	1,503	1,000	,781
Val.Mode	2,168	1,740	1,000	,803
Val.Aero	1,818	1,467	1,000	,807

Método de extracción: Análisis de Componentes principales.

1. Elementos de la diagonal Sʏ (las varianzas)
2. Cantidad de varianza explicada por las k-componentes seleccionadas.
3. Proporción explicada: (3) = (2) / (1)

La comunalidad asociada a una variable original es la proporción de variabilidad de dicha variable explicada por las k-componentes seleccionadas.

Varianza total explicada (parte 1)

	Componente	Autovalores iniciales		
		Total	% de la varianza	% acumulado
Bruta	1	11.474	63.290	63.290
	2	2.833	15.629	78.920
	3	1.269	7.002	85.922
	4	.853	4.708	90.629
	5	.589	3.250	93.880
	6	.411	2.269	96.149
	7	.282	1.558	97.707
	8	.232	1.282	98.990
	9	.127	.702	99.692
	10	5.586E-02	.308	100.000
Reescalada	1	11.474	63.290	63.290
	2	2.833	15.629	78.920
	3	1.269	7.002	85.922
	4	.853	4.708	90.629
	5	.589	3.250	93.880
	6	.411	2.269	96.149
	7	.282	1.558	97.707
	8	.232	1.282	98.990
	9	.127	.702	99.692
	10	5.586E-02	.308	100.000

1

Autovalores de la matriz de covarianzas = varianza de las c.p.

2

En este ejemplo: $\Sigma_n \lambda_i = 18.126$

3

% varianza λ_i = $(\lambda_i / \Sigma_n \lambda_i) \times 100$

4

% acumulado = Σ_{i-1} (% varianza λ_i)

Varianza total explicada (parte 2)

	Componente	Sumas de las saturaciones al cuadrado de la extracción		
		Total	% de la varianza	% acumulado
Bruta	1	11.474	63.290	63.290
	2	2.833	15.629	78.920
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			
Reescalada	1	5.678	56.784	56.784
	2	2.053	20.528	77.311
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			

Con dos componentes se explica ≈ el 79% de la variabilidad total

Selección de las componentes

Tres criterios:

a) Seleccionar un número de componentes tales que en conjunto recojan un porcentaje de variabilidad de al menos un 75% .

En nuestro ejemplo tenemos ≈ el 79% tomando las dos primeras componentes.

b) Seleccionar las componentes con λi (varianzas de las componentes) mayores que el promedio.

$$\text{promedio}(\lambda i) = 1/p * \Sigma_p \lambda i$$

b) Gráfico de Sedimentación

Magnitud del autovalor

Codo mas pronunciado, nos quedamos con dos componentes

Número del componente (o valor propio)

Matriz de componentes[a]				
	1 Bruta		2 Reescalada	
	Componente		Componente	
	1	2	1	2
Val.Preci	1,167	-,230	,870	-,171
Val.Fina	1,526	,005	,933	,003
Val.Cons	,977	-,217	,820	-,182
Val.Comb	1,464	-,271	,929	-,172
Val.Segu	-,443	,640	-,480	,693
Val.Conf	-,168	1,037	-,143	,883
Val.Capa	,369	,889	,291	,701
Val.Prest	-1,215	-,164	-,876	-,118
Val.Mode	-1,312	-,137	-,891	-,093
Val.Aero	-1,061	-,584	-,787	-,433

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

1 La componente bruta corresponde con la operación:

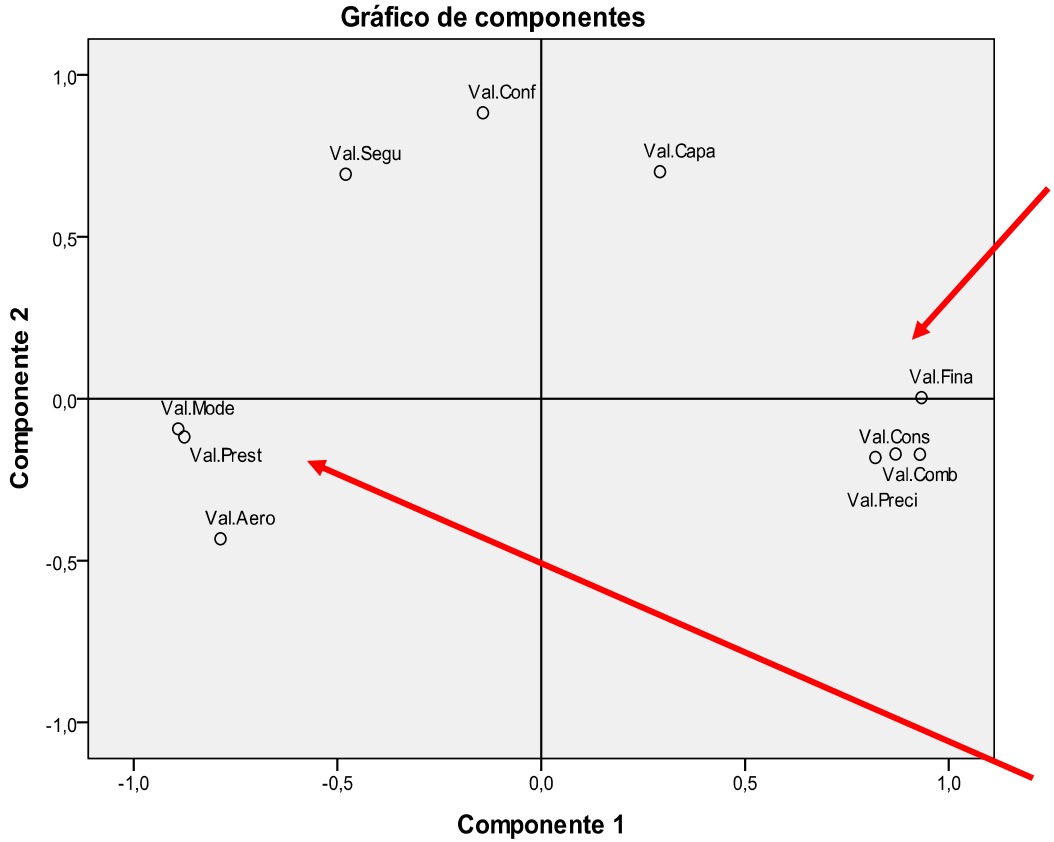
$a_{ij} = (\lambda_j)^{(1/2)} * a_{ij}$

2

La componente reescalada corresponde con la operación:

aij = componente bruta / S_{Yi}

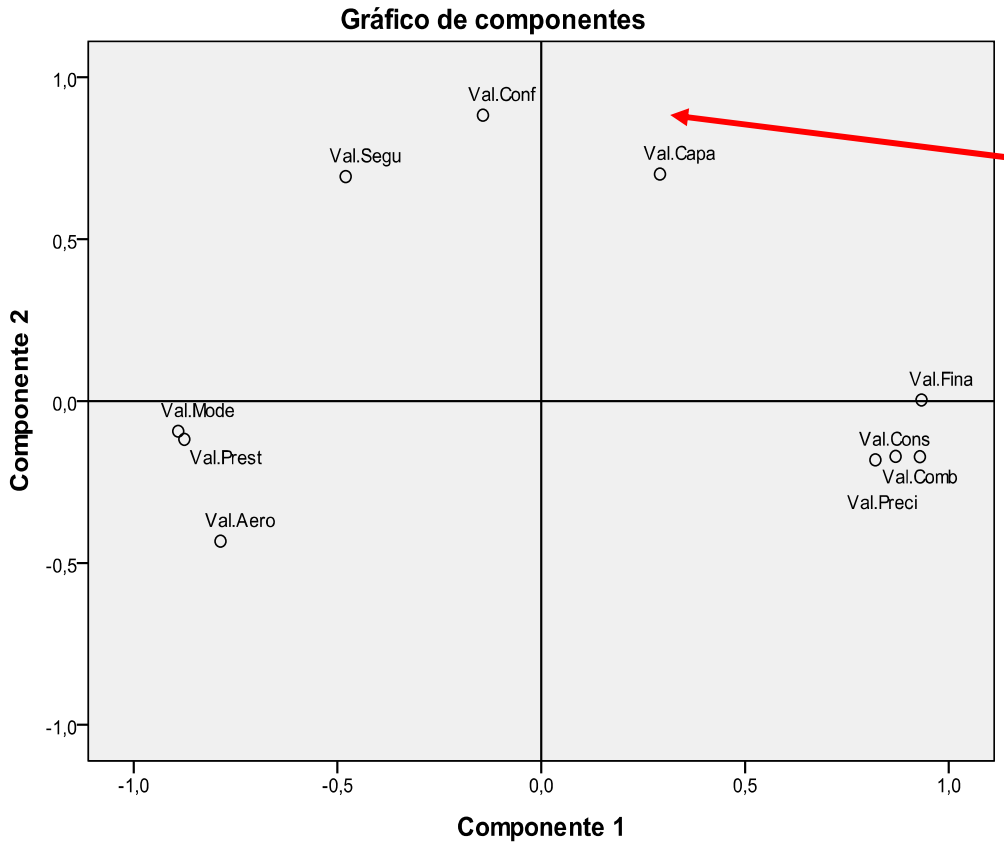
Grafico de componentes: Es un grafico de las componentes reescaladas del cuadro anterior



<u>La componente 1:</u>

Separa los compradores preocupados por la financiación (V2), combustible (V4), precio (V1) y consumo (V3)

de los compradores preocupados por la modernidad (V9), prestaciones (V8) y aerodinamica (V10)



La componente 2:

Separa los compradores preocupados por el confort (V6), capacidad (V7) y seguridad (V5)

Fin…