5. ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

Una técnica que busca describir la estructura multivariada de los datos

Análisis de componentes principales (ACP)

- **Origenes:** Karl Pearson (1901) Aplicaciones a 2- o 3-variables. Hotelling (1933)
- **Desarrollo:** Vinculado a la expansión computacional de finales del siglo XX.
- **Objetivo:** Transformar *p* variables X_1, X_2,…..X_p en *p* nuevas variables **Z_1, Z_2,…..,Z_p no correlacionadas**, denominadas **componentes principales**, que son **combinaciones lineales de las variables originales.**
- La **no correlación** de las nuevas variables puede interpretarse como que éstas "miden diferentes dimensiones" de los datos originales.

Análisis de componentes principales (ACP)

- Las nuevas variables Z_i se ordenan en función de la varianza explicada, de modo que:

$$s^2(Z_1) \geq s^2(Z_2) \geq \cdots \geq s^2(Z_p)$$

- Esto, en muchos casos, permite **reducir el número de variables originales** a un número menor de componentes principales, facilitando la interpretación.
- No obstante, debe señalarse que no siempre es posible reducir el número de variables (ej. Variables originales no correlacionadas, variables de distinta naturaleza: unidades y magnitud).

Análisis de componentes principales (ACP)

- **Origenes:** Karl Pearson (1901) Aplicaciones a 2- o 3 variables
- **Ejemplo: 2-Variables**
  Problema: Sea un proceso en el cual el test de control de calidad de la concentración de un compuesto químico en una solución fue efectuado por dos métodos diferentes.

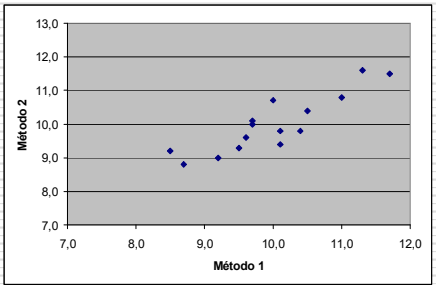
Obs. No.	Método 1 (Estándar)	Método 2 (Alternativa)
1	10,0	10,7
2	10,4	9,8
3	9,7	10,0
4	9,7	10,1
5	11,7	11,5
6	11,0	10,8
7	8,7	8,8
8	9,5	9,3
9	10,1	9,4
10	9,6	9,6
11	10,5	10,4
12	9,2	9,0
13	11,3	11,6
14	10,1	9,8
15	8,5	9,2

Hipótesis: Los métodos son intercambiables. La elección de n=15 es por conveniencia, las técnicas de control de calidad requieren normalmente un n mayor.

Análisis de componentes principales (ACP)

¿Qué podemos hacer con los datos?
Existen ∞ alternativas

Representación gráfica.
Fácil detección de "anomalías" para muestras pequeñas, así como una rápida indicación de la relación entre los dos métodos.



Análisis de componentes principales (ACP)

Regresión. ¿En qué medida es posible predecir los resultados de un método a partir del otro?
Sin embargo, el requerimiento de que esos dos métodos sean intercambiables – prediciendo en ambas direcciones – conduce a dos ecuaciones diferentes.

Mínimos cuadrados

Método 1 = f (Método 2) Minimiza variabilidad método 1
Método 2 = f (Método 1) Minimiza variabilidad método 2

Análisis de componentes principales (ACP)

Se requiere una simple ecuación predictiva que pueda ser utilizada en ambas direcciones.

Se podría invertir cualquiera de las dos ecuaciones de regresión, pero ¿cuál? Y ¿cuáles son las consecuencias técnicas de hacer esto?

La línea que cumple el requisito perseguido se denomina **Línea de Regresión Ortogonal**, que minimiza las desviaciones perpendiculares respecto a la línea. Esta línea se obtiene por el **método de componentes principales** y fue la primer aplicación del ACP (Karl Pearson, 1901).

Análisis de componentes principales (ACP)

- **Vectores y valores propios**

El análisis por componentes principales se basa en una transformación lineal de las observaciones originales, conocida en el campo del álgebra vectorial como generación de vectores y valores propios.

Análisis de componentes principales (ACP)

- Dada **S = matriz de covarianza (2 x 2), simétrica,** puede reducirse a una matriz diagonal L premultiplicando y posmultiplicando por una matriz ortonormal U.

$$U' S U = L$$

- Los elementos de la diagonal de la matriz **L** (l_1, l_2) son los **valores propios de S**. Las columnas de **U** (u_1, u_2) son los **vectores propios de S**.
- A cada valor propio corresponde un vector propio.

Análisis de componentes principales (ACP)

Propiedades

- La **suma algebraica de los valores propios**, es igual a la suma de los valores de la diagonal principal de la matriz original S.
- ➡ **es igual a la suma de las varianzas de las variables**, o sea – según definición - es la variación total.

Análisis de componentes principales (ACP)

- Los **valores propios** se determinan resolviendo la ecuación o polinomio característico de igual orden a la dimensión de la matriz dada por el determinante

$$|S - lI| = 0$$

- Para determinar los **vectores propios** se establece la condición de que estos estén **normalizados**. Esto equivale a que la suma de los cuadrados de los elementos del vector debe ser 1.

Análisis de componentes principales (ACP)

En el ejemplo bajo consideración:

$$S = \begin{bmatrix} 0.745 & 0.634 \\ 0.634 & 0.685 \end{bmatrix}$$

$$[S - lI]U = 0$$

$$\left\{ \begin{bmatrix} 0.745 & 0.634 \\ 0.634 & 0.685 \end{bmatrix} - \begin{bmatrix} l & 0 \\ 0 & l \end{bmatrix} \right\} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.745 - l & 0.634 \\ 0.634 & 0.685 - l \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad [1]$$

Análisis de componentes principales (ACP)

Operando [1] (resolviendo y sustituyendo):

$$(0.745-l)(0.685-l)u_2-(0.634)^2u_2=0$$

$$Si\ u_2\neq 0,\ se\ tiene:$$

$$\left[(0.745-l)(0.685-l)-(0.634)^2\right]u_2=0$$

$$es\ decir,\ |S-lI|=0$$

Los valores propios que satisfacen la ecuación son:

$$l_1=1.350$$

$$l_2=0.081 \qquad \sum l_i=1.431=trS$$

Análisis de componentes principales (ACP)

Igualando las ecuaciones que resultan de [1] y separando u1 y u2:

$$u_1=\frac{0.051-l}{0.111-l}u_2$$

$$u_1^2+u_2^2=1$$

e introduciendo la condición de vectores propios normalizados:

$$Para\ cada\ valor\ de\ l\ se\ tiene:$$

$$l_1=1.350 \qquad \begin{bmatrix}u_{11}\\u_{21}\end{bmatrix}=\begin{bmatrix}0.724\\0.690\end{bmatrix}$$

$$l_2=0.081 \qquad \begin{bmatrix}u_{12}\\u_{22}\end{bmatrix}=\begin{bmatrix}-0.690\\0.724\end{bmatrix}$$

$$U=[u_1u_2]=\begin{bmatrix}0.724 & -0.690\\0.690 & 0.724\end{bmatrix}$$

Análisis de componentes principales (ACP)

que es ortonormal, esto es:

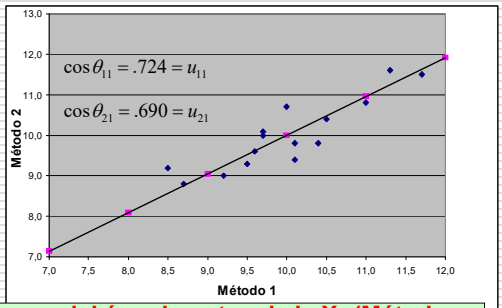
$$u_1'u_1 = 1 \quad u_2'u_2 = 1 \quad u_1'u_2 = 0$$

Además,

$$U'SU = \begin{bmatrix} .724 & .690 \\ -.690 & .724 \end{bmatrix}\begin{bmatrix} .745 & .634 \\ .634 & .685 \end{bmatrix}\begin{bmatrix} .724 & -.690 \\ .690 & .724 \end{bmatrix} = \begin{bmatrix} 1.350 & 0 \\ 0 & 0.081 \end{bmatrix} = L$$

Análisis de componentes principales (ACP)

- Geométricamente, el procedimiento descrito no es más que una rotación.
- Los elementos de los vectores propios representan las direcciones cosenos de los nuevos ejes respecto a los originales.



u_{11} es el coseno del ángulo entre el eje X_1 (Método 1) y el primero de los nuevos ejes (Z_1); u_{21} es el coseno del ángulo entre éste nuevo eje (Z_1) y el eje X_2 (Método 2).

Análisis de componentes principales (ACP)

- El nuevo eje es la línea de regresión ortogonal que estamos buscando.
- La ecuación puede resolverse para p=2 o p=3. Para valores mayores se requieren métodos iterativos (software).
- Z_2 es ortogonal a Z_1. **u_{12} es el coseno del ángulo entre el eje X_1 (Método 1) y el segundo de los nuevos ejes (Z_2); u_{22} es el coseno del ángulo entre éste nuevo eje (Z_2) y el eje X_2 (Método 2).**

Análisis de componentes principales (ACP)

Caso p variables

- El punto de partida del ACP es una matriz de datos (p variables x n observaciones)

Observaciones	X_1	X_2	...	X_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
n	x_{n1}	x_{n2}	...	x_{np}

Análisis de componentes principales (ACP)

- Seguidamente se obtiene la matriz de covarianza S

S =

s_{11}	s_{12}	...	s_{1p}
s_{21}	s_{22}	...	s_{2p}
.	.	...	.
.	.	...	.
.	.	...	.
s_{n1}	s_{n2}	...	s_{np}

$$s_{jj} = \sum_{i=1}^{n} \frac{1}{n-1}(x_{ij} - \bar{x}_j)^2$$

$$s_{jk} = \sum_{i=1}^{n} \frac{1}{n-1}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Análisis de componentes principales (ACP)

- Dada **S = matriz de covarianza (p x p), simétrica,** puede reducirse a una matriz diagonal L premultiplicando y posmultiplicando por una matriz ortonormal U.

$$U' S U = L$$

- Los elementos de la diagonal de la matriz **L** (l_1, l_2,…., l_p) son los **valores propios de S**. Las columnas de **U** (u_1, u_2, …, u_p) son los **vectores propios de S**.
- A cada valor propio corresponde un vector propio.

Análisis de componentes principales (ACP)

- Al valor propio l_i le corresponde el i-ésimo componente principal:

$$Z_i = u_{1i}X_1 + u_{2i}X_2 + \cdots + u_{pi}X_p$$

donde Z_i = componente principal
u_{ki} = elementos del i-ésimo vector propio
$Var(Z_i) = l_i$

Además se cumple que:

$$u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1$$

Interpretación de los componentes principales

- En cada nueva variable Z_i intervienen todos los valores de las variables originales X_i (i=1,2). El valor numérico de u_{ik} indicará **el grado de contribución que cada variable original aporta a la nueva variable** definida por la transformación lineal. Si u_{ik} tiene valor cero, o muy cercano a cero, indica que esa variable no influye en el valor de la nueva variable Z_i.

$$Z_1 = 0{,}724X_1 + 0{,}690X_2$$

$$Z_2 = -0{,}690X_1 + 0{,}724X_2$$

Interpretación de los componentes principales

$$Z_1 = 0{,}724 X_1 + 0{,}690 X_2$$

- Los coeficientes del primer vector, .724 y .690, son casi iguales y ambos positivos, indicando que el primer C.P., Z_1, es una media ponderada de ambas variables. Esta, por lo tanto, relacionado con la variabilidad que X_1 y X_2 tienen en común. Ya hemos visto que Z_1 define la línea de regresión ortogonal que Pearson (1901) denominaba como "la línea que mejor ajusta".

Interpretación de los componentes principales

$$Z_2 = -0{,}690 X_1 + 0{,}724 X_2$$

- Los coeficientes del segundo vector, -.690 y .724 también son casi iguales, excepto por el signo; de aquí, el segundo C.P., Z_2, representa las diferencias de medida de los dos métodos, que probablemente representa la variabilidad de los métodos y medidas. (El eje definido por Z_2 fue denominado por Pearson como "la línea que peor ajusta". Sin embargo, este término es apropiado para el vector propio correspondiente al menor valor propio, no al segundo, a menos que sólo existan dos como en este caso).

Nueva expresión de los datos

- Si se conocen los valores propios generados por la matriz de covarianza de un conjunto de datos, es posible calcular todas las constantes que forman la matriz U de transformación. Una vez encontrada esta matriz, es posible posmultiplicar la matriz original de datos (expresados como desviaciones respecto a la media, matriz **Y**) y obtener una nueva matriz de datos **Z =Y U**.

Nueva expresión de los datos

En el ejemplo bajo estudio:
Z =Y U

Z representa la nueva expresión de los datos en el eje de coordenadas Z_1 y Z_2 (C.P.)

	Método 1	Método 2			
	0,0	0,7		0,483	0,507
	0,4	-0,2		0,151	-0,421
	-0,3	0,0		-0,217	0,207
	-0,3	0,1		-0,148	0,279
	1,7	1,5		2,265	-0,088
	1,0	0,8	Z=	1,276	-0,111
Y=	-1,3	-1,2		-1,769	0,029
	-0,5	-0,7		-0,845	-0,161
	0,1	-0,6		-0,342	-0,503
	-0,4	-0,4		-0,566	-0,013
	0,5	0,4		0,638	-0,056
	-0,8	-1,0		-1,269	-0,171
	1,3	1,6		2,045	0,261
	0,1	-0,2		-0,066	-0,214
	-1,5	-0,8		-1,638	0,456

U=	0,724	-0,690
	0,690	0,724

Uso de la matriz de correlación

En muchos casos no puede utilizarse la matriz de covarianza. Dos posibles razones son:

- Las variables originales están expresadas en **unidades diferentes**. En este caso, las variables de mayor magnitud ejercen una influencia considerable sobre la forma de los C.P. puesto que el ACP tiene por objeto explicar la variabilidad.

Uso de la matriz de correlación

- Aún cuando las variables originales estén expresadas en las mismas unidades, **las varianzas pueden diferir ampliamente**. En este caso se otorga una ponderación indebida a ciertas variables.

En estos casos, pueden calcularse los valores y los vectores propios y, por lo tanto, la matriz de transformación U, empleando datos estandarizados, en cuyo caso la matriz de covarianza será la matriz de correlación. Los valores de la diagonal principal de R - la matriz de correlación - son 1, ya que las nuevas variables estandarizadas poseen varianza unitaria.

Uso de la matriz de correlación

Esto significa que en el conjunto de datos a partir del cual se generarán los componentes principales se otorga la misma importancia a todas las variables observadas. El uso de la matriz de correlación implica una ponderación de las variables originales, otorgándole a cada una la misma importancia, independientemente de los valores relativos de sus varianzas.

Uso de la matriz de correlación

- Es importante notar que la matriz de transformación U generada a partir de la matriz de correlación R será diferente de la obtenida a partir de la matriz de covarianza S. Por lo tanto, no existe una correspondencia uno a uno entre los C.P. obtenidos a partir de R y los obtenidos de S. Cuanto más heterogéneas son las varianzas, mayor será la diferencia entre los dos conjuntos de vectores.

Estas características de los valores y vectores propios determinan que el **ACP** sea **sensible a los cambios de escala.**

SELECCIÓN DEL NUMERO DE C.P.

Criterios para seleccionar el número de componentes a considerar, cuando se reduce la dimensionalidad original de p variables a k (siendo k < p)

- Cada componente principal explica una proporción de la varianza total:

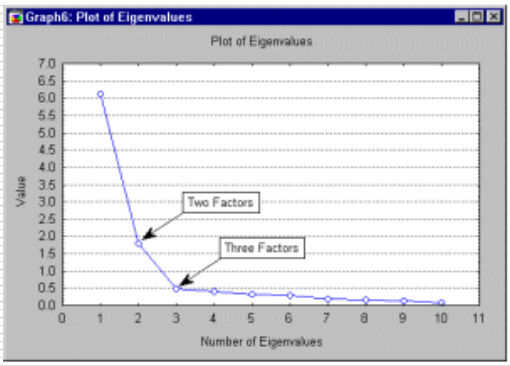
$$\frac{l_i}{tr(S)}$$

SELECCIÓN DEL NUMERO DE C.P.

- Puesto que los valores propios se ordenan en forma decreciente, es posible seleccionar los primeros k valores propios para representar los datos originales. La eficiencia de la representación por los nuevos k C.P. estará dada por la proporción de la varianza total explicada:
- Porcentaje de la varianza = $\frac{\sum_{i=1}^{k} l_i}{tr(S)} 100$
- Sin embargo, no existe un valor mágico (% de la varianza total explicada) para determinar el óptimo k.

SELECCIÓN DEL NUMERO DE C.P.

- **Método gráfico (The scree test).** Consiste en graficar los valores propios en orden decreciente. El gráfico se asemeja a la pendiente de una montaña, el término "scree" se refiere a los residuos rocosos que caen de ella y yacen sobre su base. El "scree test" propone finalizar el análisis en aquel punto en que comienza la base (residuos).
- **Método de Kaiser.** Considera solamente aquellos valores propios mayores que 1.



Correlación entre C.P. y Variables Originales

- La correlación entre cada C.P. con cada una de las variables originales puede ser útil para propósitos de diagnóstico.
- La correlación entre el i-ésimo C.P., Z_i, y la j-ésima variable original, es:

$$r_{Z_i X_j} = \frac{u_{ji}\sqrt{l_i}}{\sqrt{s_{jj}}}$$

Correlación entre C.P. y Variables Originales

- Por ejemplo, la correlación entre Z_1 y X_1 es:

$$\frac{u_{11}l_1}{\sqrt{s_{11}}} = \frac{.724\sqrt{1.431}}{\sqrt{.745}} = .974$$

- Y las correlaciones para este ejemplo son:

$$\begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cc} Z_1 & Z_2 \end{array} \\ \begin{bmatrix} .974 & -.227 \\ .969 & .248 \end{bmatrix} \end{array}$$

El primer c.p. está mucho más correlacionado con las variables originales que el segundo c.p.

Inversión del modelo de C.P.

- Otra interesante propiedad del A.C.P. es que el modelo puede invertirse, de forma tal que las variables originales pueden expresarse en función de los C.P. Dicho de otra manera, cada variable puede expresarse como una combinación lineal de los C.P.

$$X_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1p}F_p$$
$$X_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2p}F_p$$
$$\vdots$$
$$X_p = a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pp}F_p$$

APLICACIONES DEL A.C.P.

- **Ej. 2 Variables – Test de control**

Mean and standard deviation of the columns:

	Mean	Standard deviation
Método 1	10,000	0,894
Método 2	10,000	0,857

MEDIAS Y DESVIACIÓN ESTÁNDAR

Covariance matrix:

MATRIZ DE COVARIANZA (1/n-1)

	Método 1	Método 2
Método 1	0,799	0,679
Método 2	0,679	0,734

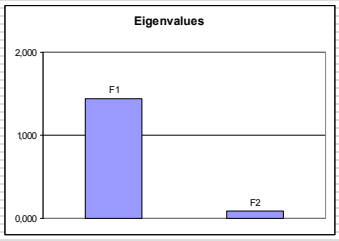
APLICACIONES DEL A.C.P.

Eigenvalues:

VALORES PROPIOS

	F1	F2
Eigenvalue	1,446	0,086
% variance	94,365	5,635
Cumulative %	94,365	100,000

Los valores propios reflejan la calidad de la proyección del espacio p-dimensional a un espacio de menor dimensión. Lo ideal es que los dos o tres primeros valores propios representen un alto porcentaje de la varianza.



APLICACIONES DEL A.C.P.

Eigenvectors:

	F1	F2
Método 1	0,724	-0,690
Método 2	0,690	0,724

VECTORES PROPIOS

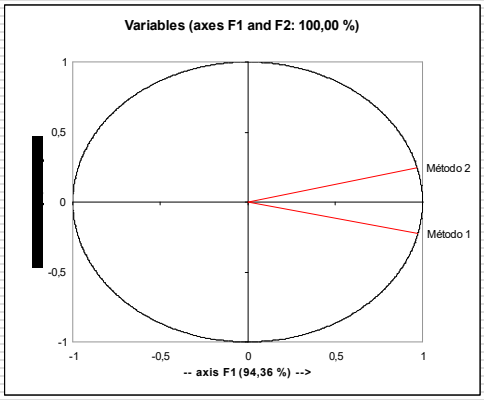
MATRIZ U
Representa las direcciones coseno de los nuevos ejes F1 y F2 o Componentes Principales

APLICACIONES DEL A.C.P.

Correlations of the variables with the factors:

	F1	F2	
Método 1	0,974	-0,227	X1
Método 2	0,969	0,248	X2

CORRELACIONES ENTRE LAS VARIABLES ORIGINALES Y LOS C.P.



APLICACIONES DEL A.C.P.

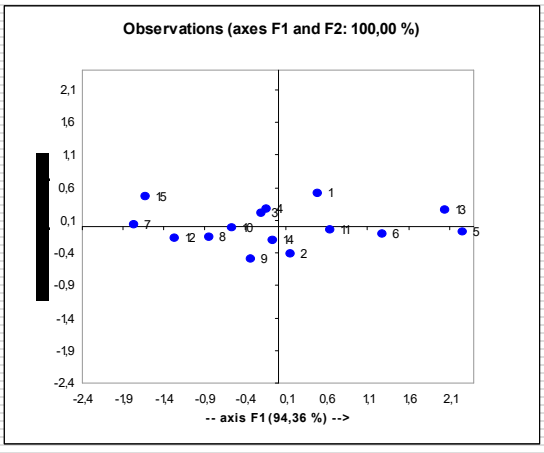
Factor scores:

	F1	F2
1	0,483	0,507
2	0,151	-0,421
3	-0,217	0,207
4	-0,148	0,279
5	2,265	-0,088
6	1,276	-0,111
7	-1,769	0,029
8	-0,845	-0,161
9	-0,342	-0,503
10	-0,566	-0,013
11	0,638	-0,056
12	-1,269	-0,171
13	2,045	0,261
14	-0,066	-0,214
15	-1,638	0,456

Z-SCORES O ESCORES FACTORIALES

COORDENADAS DE LAS OBSERVACIONES EN EL ESPACIO DE LAS NUEVAS VARIABLES O C.P.

APLICACIONES DEL A.C.P.



APLICACIONES DEL A.C.P.

Ejemplo 5.1 Medidas de los gorriones s y no-s

	X1	X2	X3	X4	X5
1	156	245	31,6	18,5	20,5
2	154	240	30,4	17,9	19,6
3	153	240	31,0	18,4	20,6
4	153	236	30,9	17,7	20,2
5	155	243	31,5	18,6	20,3
6	163	247	32,0	19,0	20,9
7	157	238	30,9	18,4	20,2
8	155	239	32,8	18,6	21,2
9	164	248	32,7	19,1	21,1
10	158	238	31,0	18,8	22,0
Sample variance	13,081	25,159	0,619	0,312	0,963

La magnitud de las variables X1 y X2 es claramente superior en comparación con las restantes variables. También lo es la varianza de estas variables. Por lo tanto, es conveniente utilizar la matriz de correlación para determinar los valores y vectores propios.

APLICACIONES DEL A.C.P.

Correlation matrix:

	X1	X2	X3	X4	X5
X1	1	**0,735**	**0,662**	**0,645**	**0,605**
X2	**0,735**	1	**0,674**	**0,769**	**0,529**
X3	**0,662**	**0,674**	1	**0,763**	**0,526**
X4	**0,645**	**0,769**	**0,763**	1	**0,607**
X5	**0,605**	**0,529**	**0,526**	**0,607**	1

In bold, significant values (except diagonal) at the level of significance alpha=0,050 (two-tailed test)

Eigenvalues:

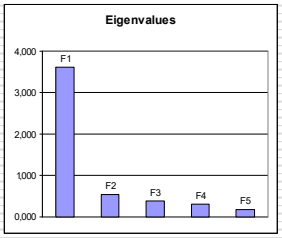
	F1	F2	F3	F4	F5
Eigenvalue	3,616	0,532	0,386	0,302	0,165
% variance	72,320	10,630	7,728	6,031	3,291
Cumulative %	72,320	82,950	90,678	96,709	100,000

El componente F1 es el más importante pues representa el 72,3% de la varianza total.

APLICACIONES DEL A.C.P.

Eigenvalues:

	F1	F2	F3	F4	F5
Eigenvalue	3,616	0,532	0,386	0,302	0,165
% variance	72,320	10,630	7,728	6,031	3,291
Cumulative %	72,320	82,950	90,678	96,709	100,000



Los dos primeros C.P. representan el 83% de la varianza total. Si se agrega un tercer C.P., se explica el 90,7% de la varianza total.

APLICACIONES DEL A.C.P.

Eigenvectors:

	F1	F2	F3	F4	F5
X1	0,452	0,051	0,690	0,420	0,374
X2	0,462	-0,300	0,341	-0,548	-0,530
X3	0,451	-0,325	-0,454	0,606	-0,343
X4	0,471	-0,185	-0,411	-0,388	0,652
X5	0,398	0,876	-0,178	-0,069	-0,192

$$Z_1 = 0{,}452X_1 + 0{,}462X_2 + 0{,}451X_3 + 0{,}471X_4 + 0{,}398X_5$$

- X_1 a X_5 son las variables originales estandarizadas, puesto que trabajamos con la matriz de correlación.
- Los coeficientes son casi iguales, por tanto Z_1 es una media ponderada del tamaño de los gorriones. Es decir, el 72,3% de la varianza de los datos está relacionada con las diferencias en tamaño.

APLICACIONES DEL A.C.P.

El segundo C.P. es:

$$Z_2 = 0{,}051X_1 - 0{,}300X_2 - 0{,}325X_3 - 0{,}185X_4 + 0{,}877X_5$$

representa un contraste entre las variables X_1,X_5 y las variables X_2, X_3, X_4.

- Z_2 representa, por lo tanto, las diferencias de forma entre los gorriones. El bajo coeficiente de X_1(longitud total) significa que el valor de esta variable no afecta el valor de Z_2.

APLICACIONES DEL A.C.P.

- Los C.P. Z_3, Z_4 y Z_5 representan otros aspectos de las diferencias de forma.

Eigenvectors:

	F1	F2	F3	F4	F5
X1	0,452	0,051	0,690	0,420	0,374
X2	0,462	-0,300	0,341	-0,548	-0,530
X3	0,451	-0,325	-0,454	0,606	-0,343
X4	0,471	-0,185	-0,411	-0,388	0,652
X5	0,398	0,876	-0,178	-0,069	-0,192

APLICACIONES DEL A.C.P.

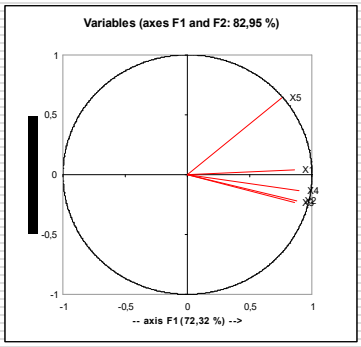
Es importante darse cuenta de que algunos programas informáticos pueden dar los componentes principales como se muestra en este ejemplo, pero con los signos de los coeficientes de las mediciones del cuerpo invertido. Por ejemplo, Z_2 puede ser mostrado como

$$Z_2 = -0{,}051X_1 + 0{,}300X_2 + 0{,}325X_3 + 0{,}185X_4 - 0{,}877X_5$$

Esto no es un error. El componente principal está midiendo exactamente el mismo aspecto de los datos, pero en la dirección opuesta.

APLICACIONES DEL A.C.P.

Correlaciones entre las variables originales y los dos primeros C.P.:



APLICACIONES DEL A.C.P.

Coordenadas de las observaciones en función de los nuevas variables o C.P.

Factor scores:

	F1	F2	F3	F4	F5
Obs1	0,064	-0,601	-0,171	-0,516	-0,549
Obs2	-2,180	-0,442	0,400	-0,645	-0,231
Obs3	-1,146	0,019	-0,676	-0,716	-0,209
Obs4	-2,311	0,172	-0,306	0,149	-0,478
Obs5	-0,295	-0,665	-0,474	-0,546	-0,244
Obs6	1,916	-0,595	0,621	0,007	0,286
Obs7	-1,050	-0,120	0,074	-0,088	0,530
Obs8	0,439	-0,164	-1,648	0,816	-0,562
Obs9	2,691	-0,782	0,368	0,465	0,058
Obs10	0,186	1,314	-0,409	-0,297	0,702
Obs11	0,371	1,138	-0,301	-0,147	0,133

APLICACIONES DEL A.C.P.

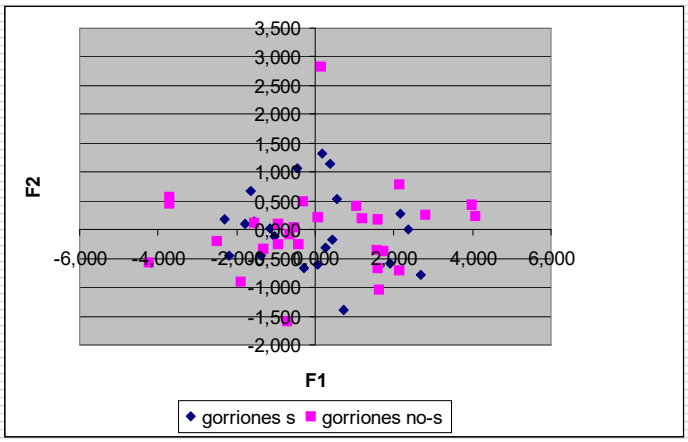
Recordemos resultados anteriores:

Ejemplo 3.1 No hay evidencia de diferencias en los valores medios.

Ejemplo 3.2 Los sobrevivientes (gorriones s) parecen haber sido menos variables que los no sobrevientes (gorriones no-s).

APLICACIONES DEL A.C.P.

Representación gráfica de las observaciones en función de los 2 primeros C.P. (EXCEL):



APLICACIONES DEL A.C.P.

Ej. 5.2 Empleo en los países europeos

GRUPO		AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
EU	Bélgica	2,6	0,2	20,8	0,8	6,3	16,9	8,7	36,9	6,8
EU	Dinamarca	5,6	0,1	20,4	0,7	6,4	14,5	9,1	36,3	7,0
EU	Francia	5,1	0,3	20,2	0,9	7,1	16,7	10,2	33,1	6,4
EU	Alemania	3,2	0,7	24,8	1,0	9,4	17,2	9,6	28,4	5,6
EU	Grecia	22,2	0,5	19,2	1,0	6,8	18,2	5,3	19,8	6,9
EU	Irlanda	13,8	0,6	19,8	1,2	7,1	17,8	8,4	25,5	5,8
EU	Italia	8,4	1,1	21,9	0,0	9,1	21,6	4,6	28,0	5,3
EU	Luxemburgo	3,3	0,1	19,6	0,7	9,9	21,2	8,7	29,6	6,8
EU	Holanda	4,2	0,1	19,2	0,7	0,6	18,5	11,5	38,3	6,8
EU	Portugal	11,5	0,5	23,6	0,7	8,2	19,8	6,3	24,6	4,8
EU	España	9,9	0,5	21,1	0,6	9,5	20,1	5,9	26,7	5,8
EU	Reino Unido	2,2	0,7	21,3	1,2	7,0	20,2	12,4	28,4	6,5
EFTA	Austria	7,4	0,3	26,9	1,2	8,5	19,1	6,7	23,3	6,4
EFTA	Finlandia	8,5	0,2	19,3	1,2	6,8	14,6	8,6	33,2	7,5
EFTA	Islandia	10,5	0,0	18,7	0,9	10,0	14,5	8,0	30,7	6,7
EFTA	Noruega	5,8	1,1	14,6	1,1	6,5	17,6	7,6	37,5	8,1
EFTA	Suecia	3,2	0,3	19,0	0,8	6,4	14,2	9,4	39,5	7,2
EFTA	Suiza	5,6	0,0	24,7	0,0	9,2	20,5	10,7	23,1	6,2
Este	Albania	55,5	19,4	0,0	0,0	3,4	3,3	15,3	0,0	3,0
Este	Bulgaria	19,0	0,0	35,0	0,0	6,7	9,4	1,5	20,9	7,5
Este	Checoeslovaquia	12,8	37,3	0,0	0,0	8,4	10,2	1,6	22,9	6,9
Este	Hungría	15,3	28,9	0,0	0,0	6,4	13,3	0,0	27,3	8,8
Este	Polonia	23,6	3,9	24,1	0,9	6,3	10,3	1,3	24,5	5,2
Este	Rumania	22,0	2,6	37,9	2,0	5,8	6,9	0,6	15,3	6,8
Este	ex URSS	18,5	0,0	28,8	0,0	10,2	7,9	0,6	25,6	8,4
Este	ex Yugoeslavia	5,0	2,2	38,7	2,2	8,1	13,8	3,1	19,1	7,8
Otro	Chipre	13,5	0,3	19,0	0,5	9,1	23,7	6,7	21,2	6,0
Otro	Gibraltar	0,0	0,0	6,8	2,0	16,9	24,5	10,8	34,0	5,0
Otro	Malta	2,6	0,6	27,9	1,5	4,6	10,2	3,9	41,6	7,2
Otro	Turquía	44,8	0,9	15,3	0,2	5,2	12,4	2,4	14,5	4,4

AGR=agricultura
MIN=minería
MAN=manufactura
PS = generación y abastecimiento agua
CON=construcción
SER=servicios
FIN=finanzas
SPS=servicios personales y sociales
TC=transporte y comunicaciones

APLICACIONES DEL A.C.P.

Matriz de correlaciones (Pearson (n-1)):

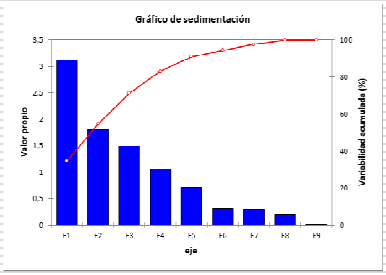
Variables	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
AGR	**1**	0,316	-0,254	-0,382	-0,349	-0,605	-0,176	-0,811	-0,487
MIN	0,316	**1**	-0,672	-0,387	-0,129	-0,407	-0,248	-0,316	0,045
MAN	-0,254	-0,672	**1**	0,388	-0,034	-0,033	-0,274	0,050	0,243
PS	-0,382	-0,387	0,388	**1**	0,165	0,155	0,094	0,238	0,105
CON	-0,349	-0,129	-0,034	0,165	**1**	0,473	-0,018	0,072	-0,055
SER	-0,605	-0,407	-0,033	0,155	0,473	**1**	0,379	0,388	-0,085
FIN	-0,176	-0,248	-0,274	0,094	-0,018	0,379	**1**	0,166	-0,391
SPS	-0,811	-0,316	0,050	0,238	0,072	0,388	0,166	**1**	0,475
TC	-0,487	0,045	0,243	0,105	-0,055	-0,085	-0,391	0,475	**1**

En general, los valores de esta matriz no son particularmente altos, lo que indica que se requerirán varios componentes principales para explicar la variación en los datos.

APLICACIONES DEL A.C.P.

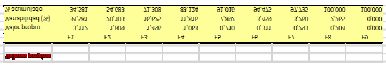
Valores propios:

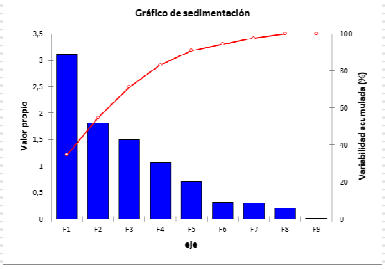
	F1	F2	F3	F4	F5	F6	F7	F8	F9
Valor propio	3,112	1,809	1,496	1,063	0,710	0,311	0,293	0,204	0,000
Variabilidad (%)	34,581	20,103	16,625	11,816	7,892	3,459	3,260	2,265	0,000
% acumulado	34,581	54,683	71,308	83,124	91,016	94,475	97,735	100,000	100,000



El último valor propio es 0 puesto que la suma de las 9 variables bajo análisis es 100% antes de la estandarización. El vector propio o C.P. correspondiente a este valor propio tiene valor 0 para todos los países, y por lo tanto tiene una varianza 0. Si cualquier combinación lineal de las variables originales en un ACP es constante, entonces esto necesariamente tendrá como resultado que uno de los valores propios sea 0.

APLICACIONES DEL A.C.P.





Este ejemplo no es tan sencillo como el anterior. El primer C.P. representa sólo el 35% de la variación en los datos, y 4 C.P. son necesarios para explicar el 83% de la variación. Es una cuestión de juicio cuántos C.P. son importantes. Se puede argumentar que sólo los primeros 4 deben ser considerados porque son aquellos con valores propios > 1. Hasta cierto punto, la elección del número de C.P. que son importantes dependerá del uso que se va a hacer de ellos. Para este ejemplo, se supondrá que un pequeño número de índices son necesarios para mostrar las principales diferencias entre los países, y por simplicidad se examinarán sólo los dos primeros C.P. que explican el 55% de la variación en los datos originales.

APLICACIONES DEL A.C.P.

Vectores propios:

	F1	F2	F3	F4	F5	F6	F7	F8	F9
AGR	0,511	0,023	-0,279	0,016	-0,024	-0,042	0,164	0,540	0,582
MIN	0,375	0,000	0,515	0,114	0,346	0,199	-0,213	-0,449	0,419
MAN	-0,246	-0,432	-0,502	0,058	-0,234	-0,031	-0,236	-0,432	0,447
PS	-0,316	-0,109	-0,294	0,023	0,854	0,206	0,061	0,155	0,030
CON	-0,222	0,242	0,072	0,783	0,062	-0,503	0,020	0,031	0,129
SER	-0,382	0,408	0,065	0,169	-0,267	0,673	-0,175	0,202	0,245
FIN	-0,131	0,553	-0,096	-0,489	0,131	-0,406	-0,458	-0,027	0,191
SPS	-0,428	-0,055	0,360	-0,317	-0,046	-0,158	0,621	-0,041	0,410
TC	-0,205	-0,517	0,413	-0,042	-0,023	-0,142	-0,492	0,502	0,061

$$Z_1 = 0{,}51(AGR) + 0{,}37(MIN) - 0{,}25(MAN) - 0{,}31(PS) - 0{,}22(CON) - 0{,}38(SER) - 0{,}13(FIN) - 0{,}42(SPS) - 0{,}21(TC)$$

Representa el contraste entre AGR y MIN versus los ocupados en otras actividades

APLICACIONES DEL A.C.P.

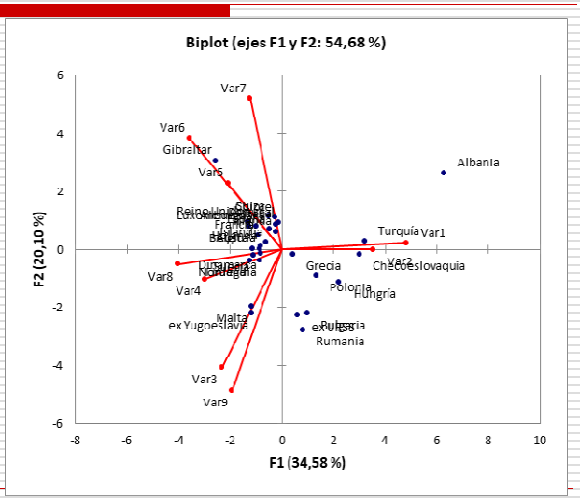
El segundo C.P. es:

$$Z_2 = 0{,}02(AGR) + 0{,}00(MIN) - 0{,}43(MAN) - 0{,}11(PS) + 0{,}24(CON) + 0{,}41(SER) + 0{,}55(FIN) - 0{,}06(SPS) - 0{,}52(TC)$$

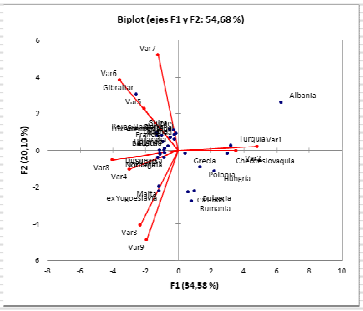
que contrasta principalmente los números de MAN y TC con los números en CON, SER y FIN.

APLICACIONES DEL A.C.P.

Agrupamiento de los países en función de los componentes principales



APLICACIONES DEL A.C.P.



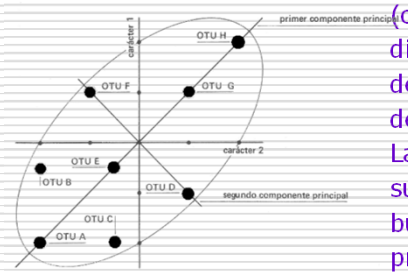
La figura es sin duda bastante significativa en términos de lo que se sabe sobre los países. La mayoría de las democracias occidentales tradicionales se agrupan con valores ligeramente negativos para F1 y positivos para F2. Gibraltar y Albania destacan por tener patrones de empleo bastante distintos, mientras que los restantes países se encuentran en una banda que va desde la antigua Yugoslavia hasta Turquía.

Análisis en modo R y en modo Q

Análisis en modo R

- Los ejemplos presentados anteriormente son ejemplos de análisis en ***modo R***. Dadas n observaciones sobre p variables se calcula la matriz de covarianza o correlación (p x p) y se determinan los valores propios y vectores propios. Estos últimos se utilizan para *representar cada vector de observaciones en términos de componentes principales.* En estos casos, **el análisis se centra en las variables (el objetivo es reducir el número de variables).**

Análisis en modo R y en modo Q



Geométricamente, la varianza es la dispersión de los puntos (observaciones) en una determinada dirección. Por lo tanto, el C.P. I se definirá en la dirección del eje mayor del hiperelipsoide (vector propio I). La varianza explicada está dada por su longitud (valor propio I). Luego se busca la dirección ortogonal al vector propio I que contenga la máxima varianza remanente (vector propio II) para definir el C.P. II, y así sucesivamente.

Análisis en modo R y en modo Q

Análisis en modo Q

- Matemáticamente es posible hacerlo en el otro sentido, esto es, obtener una matriz de covarianza o correlación (n x n), determinar los valores y vectores propios, y representar cada variable en términos de componentes principales. Este metodología se denomina análisis en ***modo Q.***
- El análisis en modo Q está diseñado para estudiar las **interrelaciones entre observaciones**. El objetivo es encontrar grupos de observaciones que sean similares entre sí en función de su composición total.

TUTORIAL XLSTAT

Datos: US Census Bureau (http://eire.census.gov/popest/states_dataset.csv). Medición de parámetros demográficos en 51 Estados de los Estados-Unidos en 2001.

State	Total Pop. I	Net Domestic Mig.	Federal/Civilian	Net Int. Migration	Period Births	Period Deaths	< 65 Pop. Est.	> 65 Pop. Est.
Alabar	4464356	-1,78	-0,02	0,69	14,41	10,28	869,21	130,79
Alaska	634892	-1,72	-0,24	2,09	15,95	4,64	941,95	58,05
Arizon	5307331	14,25	-0,03	4,29	15,88	7,77	869,54	130,46
Arkans	2692090	0,36	-0,01	1,07	14,35	10,51	861,06	138,94

Con el fin de suprimir los efectos de escala, las variables iniciales fueron convertidas en índices por 1000 habitantes.

TUTORIAL XLSTAT

- **Objetivo:** Analizar las correlaciones entre las variables e identificar Estados que se distinguen fuertemente de los demás.

  El ACP es un método muy eficaz para el análisis de datos cuantitativos (continuos o discretos) que se presentan bajo la forma de tablas de ***p* variables x *n* observaciones .**

TUTORIAL XLSTAT

El ACP permite:

- Visualizar y analizar las correlaciones entre p variables.
- Visualizar y analizar n observaciones descritas por p variables en un gráfico de dos o tres dimensiones, que preserva en lo posible la dispersión entre los datos originales.
- Construir un conjunto de p nuevas variables o C.P. no correlacionados (p<=n), que pueden luego ser reutilizados por otros métodos (por ej.: la regresión).

TUTORIAL XLSTAT

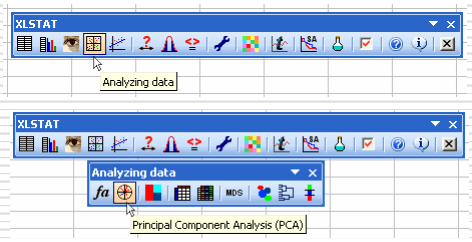
- **Límites del ACP:**
- Es un método de proyección, y, por consiguiente, la pérdida de información inducida por la proyección puede provocar interpretaciones erróneas.

TUTORIAL XLSTAT

- Activar XLSTAT-Pro
- Seleccionar en el menú XLSTAT/ Análisis de datos/ Análisis de Componentes Principales, o
- seleccionar el botón "Análisis de datos" en la barra de herramientas

y a continuación

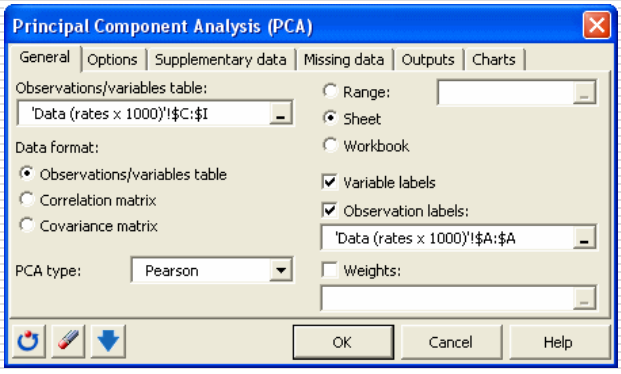
PCA.



TUTORIAL XLSTAT

En el cuadro de diálogo:

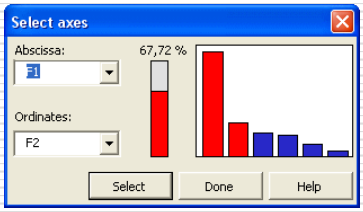
- Seleccionar los datos en la hoja Excel.

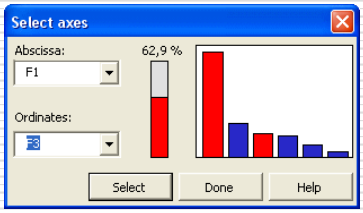


TUTORIAL XLSTAT

- El programa efectúa los cálculos luego de presionar el botón "OK".
- Seguidamente, un cuadro de diálogo presenta las opciones para la visualización de los gráficos.

  En este caso, el % de la varianza representado por los dos primeros C.P. no es particularmente alto (67.72%). Por lo tanto, para evitar una mala interpretación de los gráficos, se pide una visualización en los C.P. 1 y 3.





TUTORIAL XLSTAT

- Lo primero es analizar la **matriz de correlaciones:**

Correlation matrix:

	Net Domestic	ilian move f	Int. Migrat	Period Birth	eriod Death	65 Pop. Es	65 Pop. Est.
Net Domes	1	0.020	0.206	-0.060	-0.232	0.095	-0.095
Federal/Ci	0.020	1	-0.133	**-0.308**	**0.422**	**-0.377**	**0.377**
Net Int. Mi	0.206	-0.133	1	**0.295**	**-0.412**	0.204	-0.204
Period Birt	-0.060	**-0.308**	**0.295**	1	**-0.506**	**0.640**	**-0.640**
Period Dea	-0.232	**0.422**	**-0.412**	**-0.506**	1	**-0.779**	**0.779**
< 65 Pop.	0.095	**-0.377**	0.204	**0.640**	**-0.779**	1	**-1.000**
> 65 Pop.	-0.095	**0.377**	-0.204	**-0.640**	**0.779**	**-1.000**	1

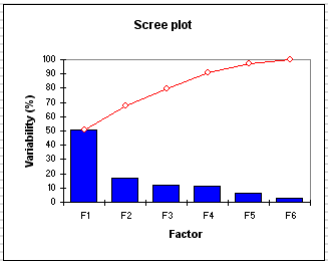
In bold, significant values (except diagonal) at the level of significance alpha=0.050 (Two-tailed test)

TUTORIAL XLSTAT

- Los índices de población > y < de 65 años de edad están perfectamente correlacionados (r = -1), por lo que las dos variables son redundantes.
- La inmigración procedente de otros estados de EE UU esta muy poco correlacionada con las restantes variables, incluso con la inmigración procedente de países extranjeros. Eso indica que las razones de inmigración son seguramente diferentes para ambos grupos de población.

TUTORIAL XLSTAT

- El siguiente cuadro y gráfico muestran los valores propios, que indican la calidad de la proyección cuando pasamos de p dimensiones (variables) a un número menor.



Eigenvalues:

	F1	F2	F3	F4	F5	F6
Eigenvalue	3.567	1.173	0.835	0.776	0.444	0.204
variance %	50.964	16.756	11.932	11.091	6.342	2.914
cumulated	50.964	67.720	79.652	90.744	97.086	100.000

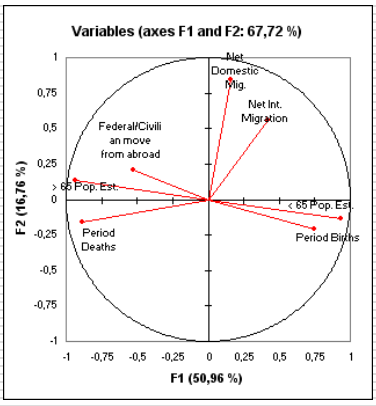
TUTORIAL XLSTAT

- **Solución ideal:** Los dos primeros C.P. explican un % elevado de la varianza, de modo que la representación en el espacio de los dos primeros ejes factoriales sea de buena calidad.
- En el ejemplo del "Tutorial", ese no es el caso, por lo que es necesario confirmar las hipótesis formuladas a partir del gráfico en los factores F1 y F2, con el gráfico en F1 y F3.
- Vemos también que el número de factores es 6, cuando las variables originales son 7. Eso se debe a que dos de las variables son redundantes. El ACP detecta automáticamente el número máximo de dimensiones "útiles", en este caso 6.

TUTORIAL XLSTAT

Círculo de correlaciones

Es el primer gráfico específico al método. Corresponde a una proyección de las variables originales en el plano de dos dimensiones determinado por los dos primeros factores. Los puntos en rojo en el gráfico representan la correlación entre variables originales y C.P. (F1 y F2).



TUTORIAL XLSTAT

Interpretación del círculo de correlaciones

Cuando dos variables están alejadas del centro del gráfico:

- Si están próximas unas de las otras, están significativamente positivamente correlacionadas (r próximo a 1).
- Si están en posición ortogonal unas respecto a las otras, están significativamente no- correlacionadas (r próximo a 0).

TUTORIAL XLSTAT

- Si están simétricamente opuestas con respecto al centro, están significativamente negativamente correlacionadas (r próximo a -1).
- Cuando las variables están próximas al centro del círculo, cualquier interpretación es arriesgada, y es necesario referirse a la matriz de correlaciones o a otros planos factoriales para interpretar los resultados.

TUTORIAL XLSTAT

- El circulo de correlaciones es también útil para interpretar la asociación de los nuevos ejes con las variables originales.
- En este caso, F1 está claramente asociado a la edad de la población y a su reemplazo; F2 está esencialmente asociado a la inmigración doméstica.
- Estas tendencias son particularmente interesantes para interpretar el gráfico de las observaciones en el espacio de las nuevas variables.

TUTORIAL XLSTAT

Tablas de los cosenos

- Para confirmar la asociación de una variable a un factor, debe consultarse la tabla de los cosenos: cuanto más elevado es el coseno (en valor absoluto), más asociada está la variable original al factor en cuestión; y cuanto más próximo a cero es el valor del coseno, la variable original está poco asociada al factor.
- En este caso, la inmigración internacional debe interpretarse valiéndose de los factores F2/F3.

TUTORIAL XLSTAT

Squared cosines of the variables:

	F1	F2	F3	F4
Net Domes	0.026	0.707	0.175	0.029
Federal/Civ	0.280	0.044	0.041	0.623
Net Int. Mi	0.174	0.317	0.463	0.017
Period Birt	0.559	0.043	0.043	0.074
Period Dea	0.780	0.026	0.002	0.002
< 65 Pop.	0.874	0.017	0.055	0.015
> 65 Pop.	0.874	0.017	0.055	0.015

correlations^2 = squared cosines

Correlations between variables and factors:

	F1	F2	F3	F4	F5	F6
Net Domes	0,161	0,841	-0,419	-0,170	0,248	0,026
Federal/Civ	-0,529	0,211	-0,203	0,789	-0,089	-0,052
Net Int. Mig	0,417	0,563	0,681	0,131	-0,122	0,121
Period Birt	0,748	-0,208	0,207	0,273	0,521	-0,100
Period Dea	-0,883	-0,162	0,043	0,050	0,257	0,351
< 65 Pop.	0,935	-0,132	-0,235	0,124	-0,106	0,162
> 65 Pop.	-0,935	0,132	0,235	-0,124	0,106	-0,162

TUTORIAL XLSTAT

- Representación de las observaciones/ individuos en 2D