1 ANALISIS DE DATOS MULTIDIMENSIONALES

1.1 INTRODUCCION

Vimos que es práctico asociar gráficos a la interpretación de los coeficientes de asociación empiricos; permiten visualizar la existencia de ligazon entre dos variables y de posibles tipologias de las observaciones, mientras que los coeficientes permiten medir el grado de relación. Pero la mayoria de los problemas envolucran más de dos variables. En el capitulo anterior, el modelo lineal permitio estudiar la relación de una variable a partir de un conjunto de variables explicativas. Veremos en este capitulo una forma de visualizar observaciones y variables para interpretar la estructura que contienen.

1.2 PLANTEAMIENTO GENERAL

En general un fenomeno se observa en varias dimensiones, lo que hace más complejo el estudio. Se busca entonces sintetizar los multiples aspectos del fenomeno en pocos valores. Es así que el objeto de un índice es reducir una realidad compleja a una sola dimensión, de manera a permitir comparaciones. Esta reducción es imposible sin deformar aspectos del fenomeno.
Sea la tabla de datos (Tabla 6.1) que contiene 6 variables socioeconomicas tomadas sobre 20 países de America Latina. Si queremos comparar los países tomando una o dos variables, se puede ordenar los países o graficarlos. Pero para las 6 variables, es más dificil hacerlo. El análisis en componentes principales permite hacerlo: en este metodo se propone un cambio de base, que permite una mejor descripción de los países y de los coeficientes de correlación entre las variables.
Si tuvieramos dos variables solamente - Esperanza de vida y tasa de mortalidad infantil - con el gráfico 8.1 tendriamos una buena herramienta para interpretar estos datos. Se observa, por ejemplo, que Bolivia tiene una alta mortalidad infantil y una baja esperanza de vida, mientras que en Costa Rica se da lo contrario; además se nota una relación lineal de pendiente negativa entre las dos variables (vimos en la tabla 6.2 del capitulo 6, que el coeficiente de correlación lineal es igual a -0.951).
Con tres o más variables, no se puede hacer tal representación gráfica, que seria en $\mathbb{R}^3$ o mayor dimensión. La idea del metodo es entonces hacer un cambio

de variables y, mediante aproximaciones, llevar a un conjunto de representaciones gráficas. Las nuevas variables -llamadas componentes principales- son índices que permiten interpretar mejor los datos.

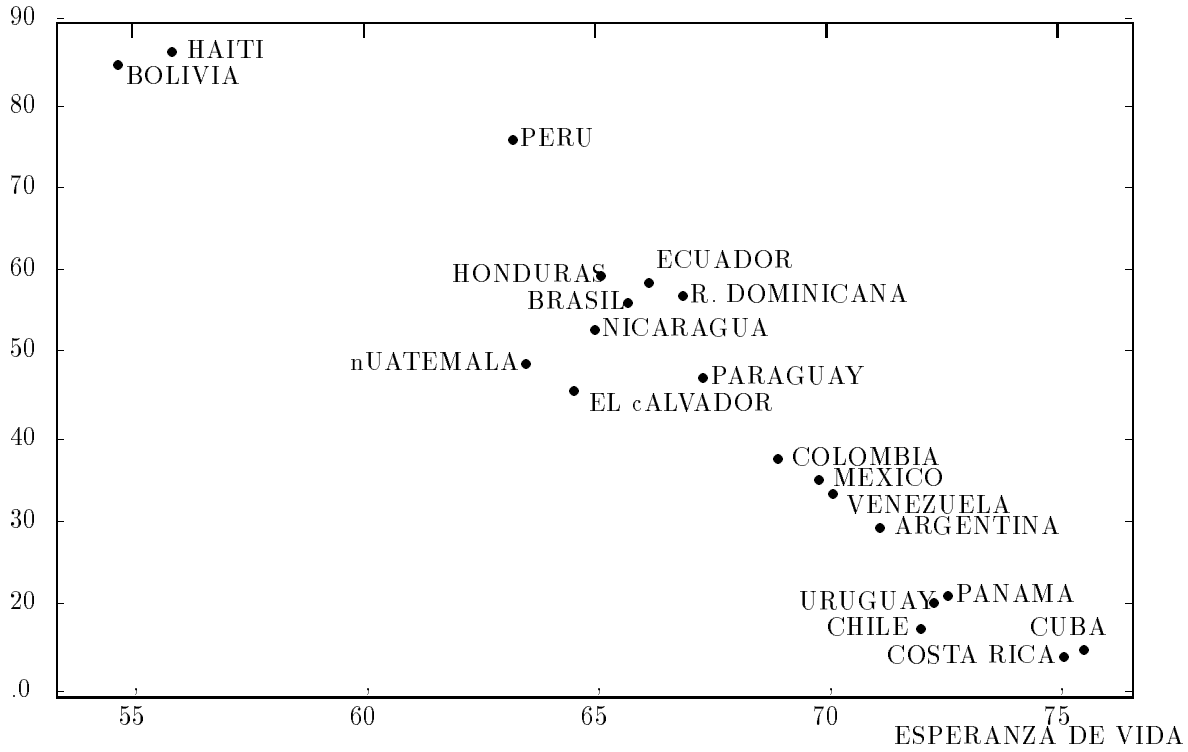


GRAFICO 8.1

1.3 EL ANALISIS EN COMPONENTES PRINCIPALES

Sea X la tabla de datos 6.1. Hay dos maneras de mirarla:

- Mirar las filas, que definen el conjunto de las observaciones

$$\mathcal{M} = \{\underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{i6} \end{pmatrix} \in I\!R^6\}$$

- Mirar las columnas, que definen el conjunto de las variables

$$\mathcal{N} = \{\underline{x}^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{20j} \end{pmatrix} \in I\!R^{20}\}$$

Se observara que si X es de rango igual a 1, entonces existe $\underline{c} \in I\!R^{20}$ y $\underline{u} \in I\!R^6$ tal que $X = \underline{c}\underline{u}^t$. En este caso, existe una recta en $I\!R^6$ que pasa por el origen, con $\underline{u}$ de vector director, a la cual pertenecen los puntos de $\mathcal{M}$; si $\underline{u}$ es unitario, las componentes c_i de $\underline{c}$ son las coordenadas de los países sobre esta recta. Sea $\|\underline{c}\|^2 = l$. Simétricamente, existe una recta en $I\!R^{20}$, que pasa por el origen, con $\underline{c}$ de vector director, a la cual pertenecen los puntos de $\mathcal{N}$. Si $\|\underline{c}\|^2 = l$, $u_j/\sqrt{l}$ son las coordenadas de las variables sobre esta recta. Si X es de rango igual a 2, entonces existe $\underline{c}_1$ y $\underline{c}_2 \in I\!R^{20}$ y $\underline{u}_1$ y $\underline{u}_2 \in I\!R^6$ tal que $X = \underline{c}_1\underline{u}_1^t + \underline{c}_2\underline{u}_2^t$. En este caso, los soportes de $\mathcal{M}$ en $I\!R^6$ y de $\mathcal{N}$ en $I\!R^{20}$ son planos. Más generalmente si X es de rango igual a r, entonces existe $\underline{c}_1, ..., \underline{c}_r \in I\!R^{20}$ y $\underline{u}_1, ..., \underline{u}_r \in I\!R^6$ tal que $X = \underline{c}_1\underline{u}_1^t + ... + \underline{c}_r\underline{u}_r^t$. En este caso, los soportes de $\mathcal{M}$ en $I\!R^6$ y de $\mathcal{N}$ en $I\!R^{20}$ son de dimensión r con estos vectores como vectores directores. El problemas es encontrar los vectores de tal descomposición.
Se distinguen las representaciones en $I\!R^6$ y en $I\!R^{20}$.

1.3.1 Representación en R^6

En este espacio los puntos son los 20 países. Para comparar dos países i e i', se considera la distancia entre las filas $\underline{x_i}$ y $\underline{x_{i'}}$ correspondientes:

$$d(i,i') = \sqrt{\sum_{j=1}^{j}(x_{ij} - x_{i'j})^2}$$

El calculo de esta distancia puede tener ciertos inconvenientes: la unidad de medición de las variables tiene un efecto, en el sentido que si multiplico por 10 una variable, por cambio de unidad, la distancia sera multiplicaado por 10 también. Se puede evitar este problema *normalizando* todas las variables, es decir tomandolas de varianza iguales a 1: si σ_j^2 es la varianza de la variable j ($\sigma_j^2 = (1/20)\sum_i(x_{ij} - \bar{x}^j)^2$), se tomara:

$$x_{ij}/\sigma_j$$

Para simplicar la notación, se supone que en la matriz X las variables son normalizadas.
Se busca entonces el vector $\underline{u} \in I\!R^6$ y el vector $\underline{c} \in I\!R^{20}$ tales que

$$X = \underline{c}\underline{u}^t + E$$

de manera que $(1/20)\sum_i^{20} \|\underline{x}_i - c_i\underline{u}\|^2$ sea mínimo con la restricción $\|\underline{u}\| = 1$. Si el rango de X es igual a 1, se obtendran los dos vectores buscados $\underline{c}$ y $\underline{u}$. La restricción $\|\underline{u}\| = 1$ se impone para tener unicidad de la solución y un vector director de la recta unitario. El criterio de optimización usado es un criterio de mínimos cuadrados que consiste a buscar la recta pasando por el origen tal que los puntos del conjunto $\mathcal{M}$ sean en promedio más cercanos a esta recta (Gráfico 8.2).

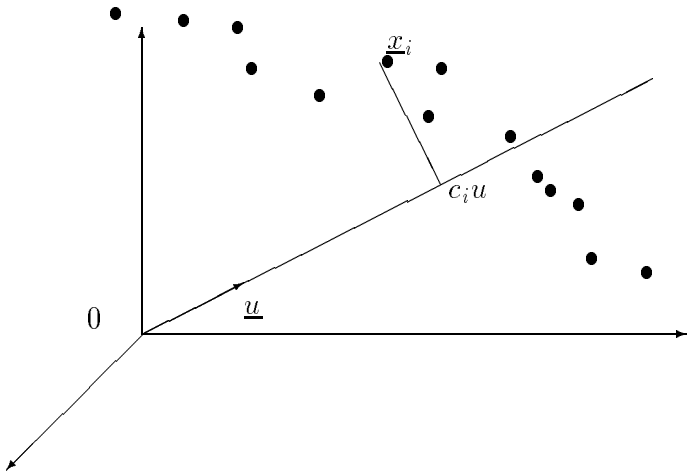


GRAFICO 8.2: Representación de las filas en $I\!R^6$

Ahora bien es un inconveniente de considerar una recta que pasa por el origen. En efecto se observa en el gráfico 8.3 que la recta H' es mejor que la recta H. Si no se impone que la recta pasa por el origen, es facil mostrar que la recta solución pasa por el punto medio $\underline{g} \in I\!R^6\}$ de $calM$: $\underline{g} = \begin{pmatrix} \bar{x}^1 \\ \bar{x}^2 \\ \\ \cdots \\ \bar{x}^6 \end{pmatrix}$, en que $\bar{x}^j$ es

la media de la variable j. En efecto, si δ es una recta pasando por el origen y δ' la recta paralela a δ pasando por $\underline{g}$ y si h_i y h_i' son las proyecciones ortogonales respectivas de $\underline{x}_i$ sobre δ y δ', entonces $\sum \|\underline{x}_i - h_i'\|^2 < \sum \|\underline{x}_i - h_i\|^2$.
De aquí se toma el origen del sistema de referencia en el punto medio, es decir $\underline{g} = \underline{0}$. Se supone entonces que en la matriz X, las columnas suman 0: $\sum_i x_{ij} = 0$ (las medias son todas nulas).

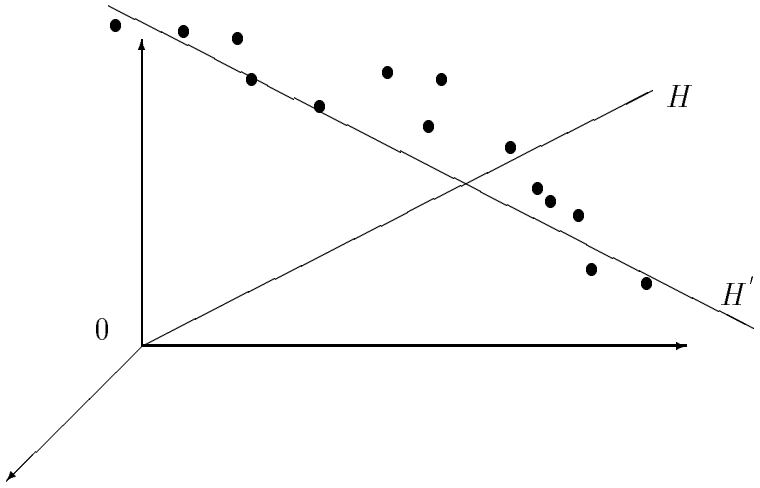


GRAFICO 8.3: Representación de las filas en $\mathbb{R}^6$

En este caso, el criterio de minimos cuadrados es equivalente a maximizar $(1/20)\sum_i \|c_i\underline{u}\|^2$ con $c_i = \underline{x}_i^t\underline{u} = \underline{u}^t\underline{x}_i$. El criterio con la restricción de norma nos da entonces:

$$Q = (1/20)\sum_i \|c_i\underline{u}\|^2 - l(\|\underline{u}\|^2 - 1)$$

$$Q = (1/20)\sum_i c_i^2 - l(\sum_j u_j^2 - 1)$$

$$Q = (1/20)\underline{u}^t(\sum_i \underline{x}_i\underline{x}_i^t)\underline{u} - l(\sum_j u_j^2 - 1)$$

$$Q = (1/20)\underline{u}^t X^t X\underline{u} - l(\sum_j u_j^2 - 1)$$

Sea $V = (1/20)X^tX = (v_{jk})$, $Q = \sum_{jk} u_j u_k v_{jk} - l(\sum u_j^2 - 1)$

$$\frac{\partial Q}{\partial u_j} = 2\sum_k v_{jk}u_k - 2lu_j = 0$$

Se deduce que $V\underline{u} = l\underline{u}$, es decir que el vector $\underline{u}$ es vector propio de la matriz $V = (1/20)X^tX$. Se observara que la matriz V es igual a la matriz de correlaciones asociada a la matriz X (Tabla 8.1) o a la matriz de covarianza cuando las variables no son normalizadas. Esta matriz es simétrica semi-definida positiva: tiene sus valores propios reales no negativos (más aún la suma de los valores propios es igual al número de variables, 6 aquí). Pero no se sabe cual de los vectores propios tomar. Observando que se busca maximizar y que $l = \sum_i c_i^2$, se concluye que hay que tomar un vector propio normalizado asociado al mayor valor propio de V. Llamamos l_1 el mayor valor propio de V, $\underline{u}_1$ el vector propio asociado y $\underline{c}_1 = X\underline{u}_1$. Si X es de rango igual a 1, l_1 es el único valor propio no nulo de V y los puntos $\underline{x}_i$ son alineados en $\mathbb{R}^6$. Si X es de rango mayor que 1, podemos repetir la descomposición a la matriz $Y = X - \underline{c}_1\underline{u}_1^t$. La matriz Y^tY tiene los mismos valores propios no nulos que X^tX salvo l_1. Luego la descomposición solución esta dada por el vector propio normalizado $\underline{u}_2$ asociado a l_2 el segundo mayor valor propio de V, y $\underline{c}_2 = X\underline{u}_2$:

$$X = \underline{c}_1\underline{u}_1 + \underline{c}_2\underline{u}_2^t + E$$

Generalizando, si $l_1 \geq l_2 \geq ... \geq l_r > 0$, $\underline{u}_1, \underline{u}_2, ..., \underline{u}_r$ los vectores propios normalizados asociados y $\underline{c}_k = X\underline{u}_k$ $k = 1, ..., r$, se puede descomponer:

$$X = \underline{c}_1\underline{u}_1^t + \underline{c}_2\underline{u}_2^t + ... + \underline{c}_r\underline{u}_r^t$$

en donde las matrices $\underline{c}_k\underline{u}_k^t$ son de rango 1 y de importancia decreciente en la reconstitución de la matriz X (Tabla 8.2).

La matriz de correlación V siendo simétrica semidefinida positiva, existe una base ortonormal de $\mathbb{R}^6$ formada de vectores propios de V. Luego, $\{\underline{u}_1, \underline{u}_2, ..., \underline{u}_r\}$ es una base ortonormal del espacio que contiene al conjunto $\mathcal{M}$.

Además se observa que los vectores $\underline{c}_k$ son vectores propios de la matriz $(1/20)XX^t$, que tiene los mismos valores propios no nulos que X^tX. En efecto, $\underline{c}_k = X\underline{u}_k$, luego

$$(1/20)X^tX\underline{u}_k = l_k\underline{u}_k$$

$$(1/20)X^t \underline{c}_k = l_k \underline{u}_k$$
$$(1/20)XX^t \underline{c}_k = l_k \underline{c}_k$$

Además $\|\underline{c}_k\|^2 = l_k$ (Se deja mostrarlo como ejercicio).

1.3.2 Representación en $\mathbb{R}^{20}$

En $\mathbb{R}^{20}$, se quiere comparar las columnas de X, que representan las variables, lo que equivale a tomar la matriz X^t en vez de X. El criterio de minimos cuadrados consiste ahora en buscar un vector $\underline{d} \in \mathbb{R}^{20}$ normalizado tal que:

$$(1/20)\sum_j^6 \|\underline{x}^j - v_j \underline{d}\|^2$$

sea mínimo.
Se tiene $v_j = d^t \underline{x}^j$ con $\|d\| = 1$.
Se obtiene que $\underline{d}$ es el vector propio normalizado de XX^t asociado al mayor valor propio l_1. Luego $\underline{d}$ es colineal al vector $\underline{c}_1$ obtenido en el estudio en $\mathbb{R}^6$: $\underline{c}_1 = \sqrt{l_1}\underline{d}$. Los vectores $\underline{u}_1$ y $\underline{v}$ son colineales también: $\underline{v} = \sqrt{l_1}\underline{u}_1$.
Interpretaremos el criterio en el caso de la representación e $\mathbb{R}^{20}$. El criterio de mínimos cuadrados es equivalente a maximizar $\Omega = (1/20)\sum_j \|v_j \underline{d}\|^2 = (1/20)\sum_j v_j^2$. Como $v_j = \underline{d}^t \underline{x}^j$ se obtiene que $\Omega = \sum_j (\underline{d}^t \underline{x}^j)^2$. Como las variables son centradas y normalizadas $\underline{d}^t x^j = Cor(\underline{d}, \underline{x}^j)$, luego el criterio usado aquí consiste en buscar una variable $\underline{d}$ de varianza igual a 1, combinación lineal de las variables $\underline{x}^j$ de tal forma que

$$\sum_j cor^2(\underline{d}, \underline{x}^j)$$

sea máxima. De hecho vimos en el capítulo 6 que el coeficiente de correlación permite comparar dos variables. Muestre como ejercicio que si dos variables son centradas y normalizadas entonces el coeficiente de correlación es igual al coseno del angulo que forman en $\mathbb{R}^{20}$ (Gráfico 8.4).
Además como los vectores $\underline{d}_k$ forman una base ortonormal, se deduce que las nuevas variables, que son las componentes principales no son correlacionadas entre si.

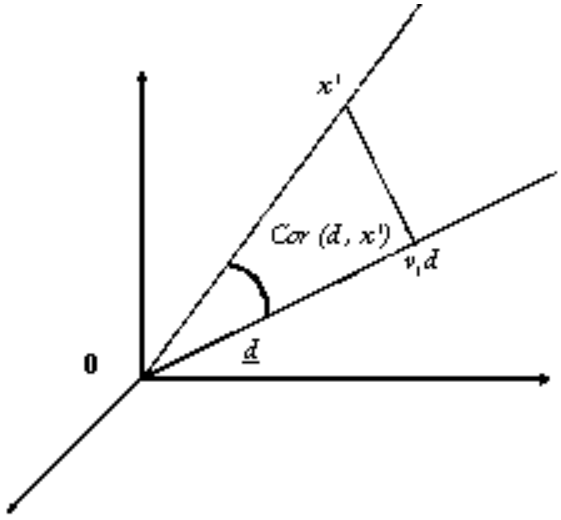


GRAFICO 8.4: Representación de las variables en $\mathbb{R}^{20}$

1.3.3 Interpretación

Veamos como usar estos resultados para interpretar el contenido de la tabla 6.1 (Se centra y normaliza los datos).
$\{\underline{u}_1, \underline{u}_2, ..., \underline{u}_r\}$ forma una base ortonormal de espacio que contiene al conjunto $\mathcal{M}$ de los países. Los ejes definidos por estos vectores se llaman **ejes principales**. Las coordenadas de los países sobre estos ejes son dadas por los vectores $\underline{c}_k$, llamados **componentes principales**, se habla también de **factores**. Si nos limitamos a tomar el primer eje principal definido por $\underline{u}_1$ (tabla 8.2), se obtiene una representación unidimensional de los países; es la mejor representación unidimensional, en el sentido que deforma menos las distancias mutuales entre los países. Aún si es una representación aproximada, tiene la ventaja de permitir una interpretación mucho más simple que la representación original. Las coordenadas de los países sobre este eje constituyen la primera componente principal $\underline{c}_1$ (Tabla 8.3). El valor más elevado lo tiene CUBA y el más bajo HAITI. Observando que

$$\underline{c}_1 = 0.3810\times\underline{x}^1 + -0.4364\times\underline{x}^2 + -0.2361\times\underline{x}^3 + 0.4590\times\underline{x}^4 + -0.4531\times\underline{x}^5 + -0.4389\times\underline{x}^6$$

se ve que la primera componente principal es una combinación lineal de las variables iniciales con algunos coeficientes mayores que otros y algunos positivos y otros negativos. EL PORCENTAJE DE POBLACION URBANA

y LA ESPERANZA DE VIDA tienen un coeficiente positivo, mientras que los otros son negativos. Lo que permite de interpretar la primera componente principal como un índice demografico, que crece con la calidad. Este índice es más manipulable que las seis variables originales. Ahora bien que cantidad de la información contenida en la tabla X perdimos o conservamos en el índice. En la decomposición: $\underline{x}_i = c_i \underline{u}_1 + \underline{e}_i$, $\underline{e}_i$ representa el error de representación de $\underline{x}_i$ sobre el primer eje principal. El valor propio $l_1 = \sum_i c_i^2$ mide la varianza de la componente principal $\underline{c}_1$ y $TrazaV - l_1 = \sum_{k=2}^{r} l_k = 6 - \_1$ mide el error global de la representación sobre el primer eje principal. Como $TrazaV = (1/20) \sum_i \|\underline{x}_i\|^2$ representa la varianza total en $\mathbb{R}^6$, se usa un índice de calidad de la representación de $\underline{c}_1$ con el porcentaje de varianza reproducida por $\underline{c}_1$:

$$100 \frac{l_1}{TrazaV}$$

que aquí vale 69.24%. Se puede considerar 2, 3 o más ejes principales para tener una mejor representación. Por ejemplo, con los dos primeros ejes principales se puede vizualisar los países (Gráfico 8.5) en un sistema cartesiano. En este gráfico cada país i tiene por coordenadas (c_{1i}, c_{2i}) y como los ejes son ortogonales, la varianza reproducida por el plano es igual a

$$100 \frac{l_1 + l_2}{TrazaV}$$

que aquí vale 88.53%.
Se nota en la tabla 8.4 que la representación con 4 ejes principales contiene casi integralmente los países (99.23%). En el gráfico de los dos primeros ejes principales (Gráfico 8.5) se proyectaron además los ejes iniciales, lo que permite explicar las diferencias y semejanzas entre los países. Es así que ARGENTINA y GUATEMALA difieren más por las variables % POBLACION URBANA, TASA NATALIDAD y FECUNDIDAD, que las variables de MORTALIDAD y ESPERANZA DE VIDA. Mientras que PANAMA y HAITI difieren más por la MORTALIDAD.
De la misma manera que se hizo una representación plana aproximada de la representación en $\mathbb{R}^6$, se hace una representación aproximada de las variables en $\mathbb{R}^{20}$, considerando las proyecciones de las variables $\underline{x}^j$ sobre los vectores $\underline{d}_1$ y $\underline{d}_2$ (Gráfico 8.6). Dado que las variables $\underline{x}^j$ y $\underline{d}_1$ y $\underline{d}_2$ son de varianza igual a 1, la proyección de $\underline{x}^j$ sobre $\underline{d}_1$ ($\underline{d}_2$) es igual al coeficiente de correlación entre

$\underline{x}^j$ y $\underline{c}_1$ ($\underline{c}_2$) (Tabla 8.4). Este gráfico permite entonces interpretar las componentes principales. Se observa que la primera componente principal tiene una correlación igual a 0.935 con la ESPERANZA DE VIDA, pero solamente -0.481 con la TASA DE MORTALIDAD, mientras que la segunda componente principal tiene una correlación igual a -0.267 con la ESPERANZA DE VIDA y 0.815 con la TASA DE MORTALIDAD.
Como las variables $\underline{x}^j$ tienen una varianza igual a 1, sus proyecciones en el plano caen al interior de un circulo de centro 0 y de radio 1. Si la proyección de la variable $\underline{x}^j$ es sobre la circunferencia del circulo, significa que x^j pertenece a este plano, es decir que $\underline{x}^j$ puede ser reproducida a partir de $\underline{c}_1$ y $\underline{c}_2$. La distancia de la proyección de una variable al origen mide la calidad de representación de la variable en el plano principal. Más aún es igual al coeficiente de correlación múltiple entre la variable con respecto a $\underline{c}_1, \underline{c}_2$ (Se deja como ejercicio la demostración). Aquí, las seis variables son bastante bien representada en el plano principal.
Como los cosenos de los angulos son iguales al los coeficientes de correlación, se tiene también una vizualisación, aproximada, de la matriz de correlaciones (Tabla 8.1). FECUNDIDAD y TASA DE NATALIDAD hacen un angulo pequeño, son altamente correlacionados (0.972), ESPERANZA DE VIDA y MORTALIDAD INFANTIL, que forman un angulo vecino de π, son altamente correlacionados negativamente (-0.951) y TASA DE MORTALIDAD y TASA DE NATALIDAD, que son casi ortogonal, son muy poco correlacionados (0.101).
Se puede completar el estudio haciendo representaciones planas con otros pares de ejes principales y las componentes principales correspondientes.

VARIABLES	1	2	3	4	5	6
1 % POB. URBANA	1.0	-.739	-.179	.588	-.735	-.532
2 TASA .NATALIDAD	-.739	1.0	.101	-.723	.972	.682
3 TASA MORTALIDAD	-.179	.101	1.0	-.609	.262	.533
4 ESPERANZA VIDA	.588	-.723	-.609	1.0	-.769	-.951
5 FECUNDIDAD	-.735	.972	.262	-.769	1.0	.709
6 MORTAL. INFANTIL	-.532	.682	.533	-.951	.709	1.0

TABLA 8.1: Matriz de correlaciones

	MEDIA	D. TIPICA	$\underline{u}_1$	$\underline{u}_2$	$\underline{u}_3$	$\underline{u}_4$
VALORES PROPIOS			4.15	1.16	0.41	0.24
% POB. URBANA	62.87	17.26	0.3810	0.3203	0.7699	0.3797
TASA NATALIDA	28.86	6.64	-0.4364	-0.3742	0.1920	0.3201
TASA MORTALIDAD	7.11	1.85	-0.2361	0.7567	-0.3904	0.4068
ESPERANZA VIDA	67.11	5.52	0.4590	-0.2479	-0.2093	0.2282
FECUNDIDAD	3.61	0.96	-0.4531	-0.2488	0.0859	0.5245
MORTAL.INFANTIL	44.54	22.28	-0.4389	0.2405	0.3779	0.5102

TABLA 8.2: Tres primeros vectores propios normalizados de la matriz de correlación

	$\underline{u}_1$	$\underline{u}_2$	$\underline{u}_3$	$\underline{u}_4$
VALORES PROPIOS	4.15	1.16	0.41	0.24
ARGENTINA	1.9029	1.3903	-.0092	0.5067
BOLIVIA	-3.1987	1.1181	.4426	-.4150
BRASIL	.2766	.8794	.6930	-.2980
COLOMBIA	1.1429	-.1561	.2443	-.3948
COSTA RICA	1.8937	-1.9713	-.6418	-.3264
CHILE	2.3727	.2178	.2529	.3686
ECUADOR	-.7182	-.1958	.1244	-.2654
EL SALVADOR	-1.1377	-.5743	-.5380	-.1134
GUATEMALA	-2.3926	-.9928	-.3855	.8457
HAITI	-4.0755	1.6465	-1.0408	-.1406
HONDURAS	-2.0627	-.8537	-.1854	.2504
MEXICO	1.0889	-.5733	.4780	-.1113
NICARAGUA	-1.8157	-.9593	.6089	.9066
PANAMA	1.5675	-1.0166	-.7486	-.4131
PARAGUAY	-.8912	-.9583	-.3161	.0197
PERU	-.8602	.8557	.9236	-.6187
REP.DOMINICANA	-.0239	-.1564	.2586	-.7430
URUGUAY	2.3890	2.2318	-.7393	.6895
VENEZUELA	1.3812	-.3757	1.2787	.5232
CUBA	3.1611	.4441	-.7003	-.2704

TABLA 8.3: Tres primeras componentes principales

	MEDIA	D. TIPICA	FACTOR .	FACTOR 2	FACTOR 3	FACTOR 4
VALORES PROPIOS			4.15	..16	'41	0.24
% ACUMULADO DE LA VARIABILIDAD			69.24	88.53	95.22	99.23
% POB. URBANA	62.87	.7.26	0.776	0.345	0.493	0.186
TASA NATALIDA	28.86	6.64	-0.889	-0.403	0.123	0.156
TASA MORTALIDAD	7.11	..85	-0.481	0.815	-0.250	0.199
ESPERANZA VIDA	67.11	5.52	0.935	-0.267	-0.134	0.111
FECUNDIDAD	3.61	0.96	-0.923	-0.268	0.055	0.256
MORTAL.INFANTIL	44.54	22.28	-0.894	0.259	0.242	-0.249

TABLA 8.4: Coordenadas de las variables sobre los 4 primeros factores (r_{jk})

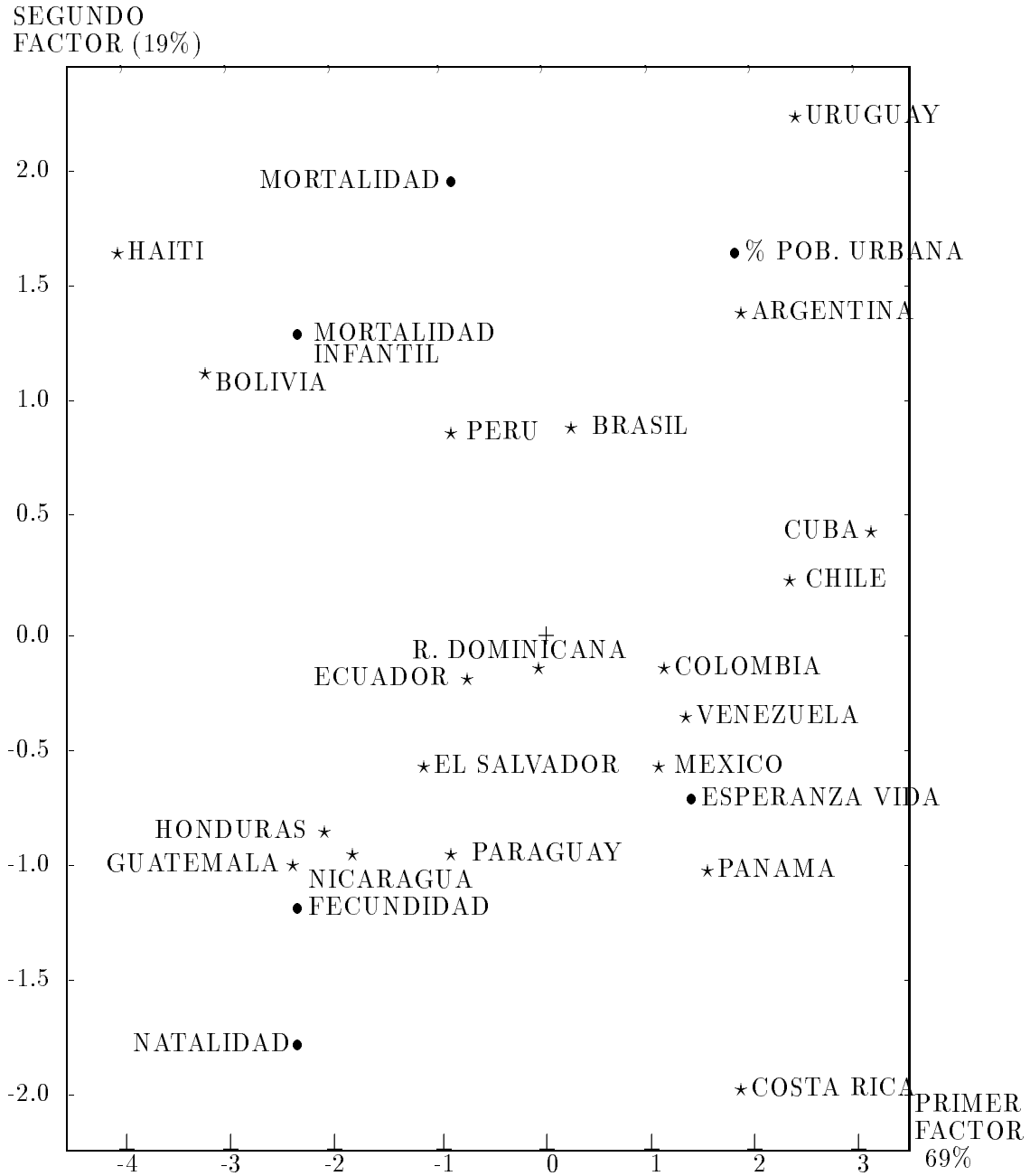


GRAFICO 8.5: Primer plano principal

GRAFICO 8.6: Círculo de correlaciones

1.3.4 Puntos suplementarios

Es interesante de representar a posteriori algunas observaciones o variables que no fueron incluídas en la matriz X originalmente. Sea un país x_o, su proyección sobre el eje principal k es igual a $\underline{x}_o^t \underline{u}_k$. Para una nueva variable $\underline{z}$, su proyeción sobre la componente principal k es igual a $Cor(\underline{z}, \underline{c}_k)$.
Consideramos, por ejemplo, dos países africanos -TUNEZ y EGIPTO- (Tabla 8.5), la coordenada de TUNEZ en el plano son (F_1, F_2) con
$F_1 = 0.3810 \times (-0.514) - 0.4364 \times 0.021 - 0.2361 \times (-0.059) + 0.4590 \times (-0.074) +$
$-0.4531 \times 0.094 + -0.4389 \times 0.155 = -0.335$
$F_2 = 0.3203 \times (-0.514) - 0.3742 \times 0.021 + 0.7567 \times (-0.059) - 0.2479 \times (-0.074) +$
$-0.2488 \times 0.094 + 0.2405 \times 0.155 = -0.18$
Para EGIPTO, se obtiene de la tabla 8.5: $F_1 = -0.18$ y $F_2 = 0.96$. Si se ubican estos dos países en el gráfico 8.5, encontramos TUNEZ cercano de R. DOMINICANA y EGIPTO cercano de BOLIVIA.

Consideramos ahora cuatro nuevas variables cuyos coeficientes de correlación con las dos primeras componentes principales son dados en la tabla 8.6. Las variables GASTO MILITAR y GASTO EN EDUCACION son muy poco correlacionados con estas componentes principales, se podria prever que un modelo lineal de estas variables sobre las seis variables originales no seria bueno. No es el caso de las dos otras variables suplementarias.

	TUNEZ		EGIPTO			
	x	$(x-\bar{x})/\sigma$	x	$(x-\bar{x})/\sigma$	$\bar{x}$	σ
% POB. URBANA	54	-0.514	47	-0.92	62.87	17.26
TASA NATALIDA	29	0.021	33	0.62	28.86	6.64
TASA MORTALIDAD	7	-0.059	10	1.56	7.11	1.85
ESPERANZA VIDA	66.7	-0.074	60.3	-1.23	67.11	5.52
FECUNDIDAD	3.7	0.094	4.3	0.72	3.61	0.96
MORTAL.INFANTIL	48.0	0.155	61.0	0.74	44.54	22.28

TABLA 8.5: Valores de las variables para TUNEZ y EGIPTO

	FACTOR 1	FACTOR 2
PNB	0.814	0.130
GASTO EN EDUCACION	-0.140	0.163
GASTO MILITAR	-0.378	-0.061
ALFABETISMO	0.839	0.021

TABLA 8.6: Coeficientes de correlación

1.4 EJERCICOS

1. Sea X la tabla siguiente: $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$

Consideremos los seis vectores $\mathcal{M} = \{\underline{x}_1, ..., \underline{x}_6\}$ de $\mathbb{R}^3$ dotado de la métrica euclidiana usual cuyas componentes están dadas por las filas de la matriz X.
a) Muestre que la nube de los 6 puntos en $\mathbb{R}^3$ está centrada en el origen. Calcule $V = (1/6)\sum_i \underline{x}_i \underline{x}_i^t$.

b) Calcule I_0, el momento de inercia de $\mathcal{N}$ con respecto al origen. Compare con $TrazaV$.
c) Determine los diferentes valores propios de V.
d) Dé el vector propio asociado al valor propio nulo de V.
e) Determine dos vectores propios ortonormales de V asociado con los valores propios no nulos de V.
2. Se consideran V_1, V_2, V_3 y V_4, cuatro variables obtenidas sobre 20 observaciones repartidas en 3 clases (A, B y C) (Tabla 8.7).
a) Los resultados del análisis en componentes principales efectuado sobre las variables V_1, V_2 y V_3 con la matriz de correlaciones (tabla 8.9) están dados en el gráfico 1 y la tabla 8.8. Justifique la calidad de la representación en el plano y comente el gráfico 8.7.
b) A partir de la tabla 8.8, dibuje y comente el círculo de correlaciones.
c) En la tabla 8.10, se dan las correlaciones entre las dos componentes principales y la variable V_4. Represente gráficamente V_4 en el círculo de correlaciones.
d) Se quiere efectuar la regresión múltiple de V_4 sobre V_1, V_2 y V_3. ¿Qué problema numérico se va a presentar?
e) Deduzca de la tabla 4 el coeficiente de correlación múltiple de la regresión de V_4 sobre V_1, V_2 y V_3.
f) Deduzca de la tabla 8.10 los coeficientes de la regresión de V_4 sobre las dos componentes principales (la media de V_4 es 242.5 y la desviación tipica es 57.73).

CLASE	V_1	V_2	V_3	V_4	CLASE	V_1	V_2	V_3	V_4
C	45	25	30	160	C	60	27	13	350
C	40	30	30	200	B	38	37	25	240
C	32	32	36	210	B	35	38	27	220
C	35	28	37	250	B	22	38	40	180
C	50	33	17	260	A	18	33	49	190
B	55	45	0	300	B	15	39	46	185
B	58	35	7	320	A	20	40	40	300
C	62	28	10	310	A	25	35	40	220
B	48	32	20	280	A	22	33	45	225
B	52	34	14	300	C	32	26	42	150

TABLA 8.7: Tabla de datos

	MEDIA	DESVIACION TIPICA	FACTOR 1	FACTOR 2
VALORES PROPIOS	1.956	1.044		
% ACUMULADOS DE LOS VALORES PROPIOS	65.20	100.00		
V_1	38.20	14.98	0.997	0.076
V_2	33.40	5.18	-0.189	-0.982
V_3	28.40	14.50	-0.962	0.273

TABLA 8.8: Correlaciones de las variables con las Componentes Principales

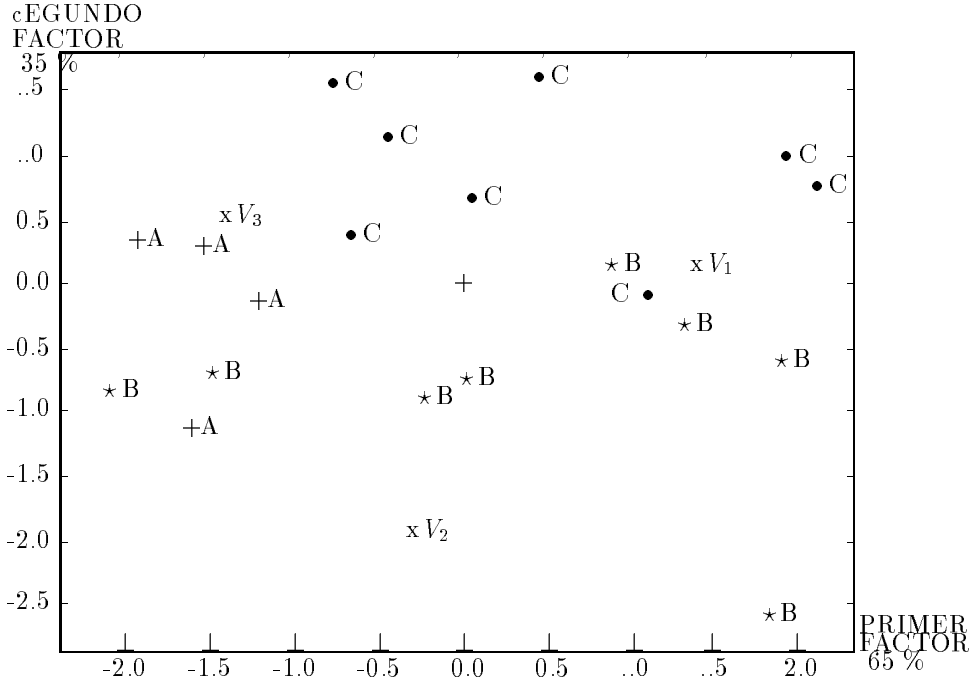


GRAFICO 8.7

	V1	V2	V3
V_1	1.00	-.26	-.94
V_2	-.26	1.00	-.09
V_3	-.94	-.09	1.00

TABLA 8.9: Matriz de correlaciones

	C.P. 1	C.P. 2	V4
C.P. 1	1.00	-.00	.69
C.P. 2	-.00	1.00	-.29
V4	.69	-.29	1.00

TABLA 8.10: Matriz de correlaciones

3. Sea $\mathcal{M} = \{\underline{x}_1, ..., \underline{x}_n\}$ un conjunto de n puntos de $\mathbb{R}^p$. Cada punto $\underline{x}_i$ tiene un peso p_i, con $p_i > 0, \sum p_i = 1$. Se supone que el centro de gravedad de $\mathcal{M}$

es $\underline{g} = 0$ y que la matriz de varianzas-covarianzas asociadas es $V = X^t D_p X$ de rango p con $D_p = diag(p_i)$. Sea $I\!R^p = W_1 \bigoplus W_2$ y sean P_1 y P_2 los proyectores ortogonales sobre W_1 y W_2 respectivamente.
a) Dé las matrices de varianzas-covarianzas V_1 y V_2 de los conjuntos $\mathcal{M}_1 = \{P_1\underline{x}_1, ..., P_1\underline{x}_n\}$ y $\mathcal{M}_2 = \{P_2\underline{x}_1, ..., P_2\underline{x}_n\}$.
b) Muestre que $V = V_1 + V_2 \Longleftrightarrow W_1 \perp_{V^{-1}} W_2$
c) Pruebe que: $[W_2 \perp_{V^{-1}} W_1] \Longleftrightarrow [V\underline{u} = l\underline{u} \Rightarrow \underline{u} \in W_1 \bigcup W_2]$
4. Sean E y F dos espacios vectoriales de dimensiones respectivas p y n. Se tiene en E y F las métricas euclidianas usuales. Sea S una aplicación lineal de E en F tal que si $\underline{y}_1 = S(\underline{x}_1)$ e $\underline{y}_2 = S(\underline{x}_2)$, entonces $\|\underline{y}_1 - \underline{y}_2\|^2 = \|\underline{x}_1 - \underline{x}_2\|^2$ para todo $\underline{x}_1, \underline{x}_2 \in E$.
a) Dar la relación que cumple S.
b) Sea $E = E_1 \oplus E_2$ con E_2 suplemento ortogonal de E_1. Sea A el proyector ortogonal sobre E_1 y S la aplicación de simetría respecto de E_2: $\underline{y} = S(\underline{x}) = -\underline{x}_1 + \underline{x}_2$.
Dar la expresión de S en función de A y mostrar que S es un isomorfismo.
c) Mostrar que S es simétrica y ortogonal.
d) En el caso de que E_2 tiene dimensión 1, se considera una nube $\mathcal{M}$ de n puntos en E y V la matriz de covarianza asociada. Se supone que hay simetría con respecto a E_2 entre los puntos de $calM$ (si $\underline{x} \in \mathcal{M}$, entonces $S(\underline{x}) \in \mathcal{M}$).
Muestre que E_2 es un eje principal de la nube $\mathcal{M}$ en E.
5. Consideremos el espacio euclidiano $I\!R^p$ dotado de una métrica euclidiana M, y un conjunto de puntos $\mathcal{M} = \{\underline{x}_i : i = 1, 2, ..., n\}$ de $I\!R^p$.
Cada punto $\underline{x}_i$ está dotado de una masa $m_i > 0$, con $\sum m_i = 1$ y suponemos que el centro de gravedad de $\mathcal{M}$ está en el origen ($\underline{g} = \sum_i m_i \underline{x}_i = \underline{0}$) y se define $V = \sum_i m_i \underline{x}_i \underline{x}_i^t$. $I\!R^p$ está descompuesto en una suma directa de dos s.e.v. M-ortogonales: Δ_u, generado por un vector u de $I\!R^p$ pasando por el origen; y el hiperplano $H = \Delta_u^{\perp}$ M-ortogonal a Δ_u pasando por el origen: $I\!R^p = \Delta_u \bigoplus H$.
a) Exprese el momento de inercia I_0 del conjunto $\mathcal{M}$ con respecto al origen en función de M.
b) Deduzca que $I_0 = tr(VM)$.
c) Muestre que $I_H = \underline{u}^t MVM\underline{u}$ donde I_H es el momento de inercia de $\mathcal{M}$ con respecto a H.

6. EXAMEN DE PRIMAVERA 1994.

PARTE 1

Se considera 6 mediciones hechas sobre 23 peces. Se presenta los resultados de un análisis en componentes principales sobre estos datos.

a) Interprete los porcentajes de los valores propios (Tabla 8.11).

b) Interprete el gráfico 8.8: ?' Que tamaño y forma tienen los peces 1, 5, 8 y 11?

c) Gráfique el circulo de correlación a partir de la tabla 8.11 y comente.

d) Usando la tabla 8.11 dé las expresiones de las primeras componentes principales C_1 y C_2 en función de las 6 mediciones. Interprételas.

e) Usando la matriz de correlaciones (tabla 8.12), ubique las variables suplementarias PESO y RADIOACTIVIDAD en el círculo de de correlaciones.

f) Se quiere hacer el modelo lineal: $PESO = \beta_o + \beta_1 \underline{c}_1 + \beta_2 \underline{c}_2$, en donde $\underline{c}_1$ y $\underline{c}_2$ son las dos primeras componentes principales. Dé el coeficiente de correlación múltiple R^2.

PARTE 2

a) Se quiere hacer el modelo lineal: $RADIOACTIVIDAD\ Y = \beta_o + \beta_1 \underline{c}_1 + \beta_2 \underline{c}_2$. Sea X la matriz (23x3) asociado a este modelo lineal. Calcule la matriz $(X^t X)^{-1}$.

b) Calcule $X^t Y$, en donde Y es el vector a explicar del modelo lineal.

c) Dé los estimadores de mínimos cuadrados de β_o, β_1 y β_2.

d) Dé el coeficiente de correlación múltiple R^2. Deduzca el estimador insesgado de la varianza σ^2 de los errores y la estimación de la varianza de los estimadores.

e) Muestre que los estimadores de β_1 y β_2 son no correlacionados. Haciendo el supuesto de normalidad, encuentre intervalos de confianza de nivel 95% para β_1 y β_2.

f) Efectúe los tests de hipótesis $H_o : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$.

PARTE 3

a) Los 23 peces estan divididos en tres acuarios. Se busca si el acuario tiene un efecto sobre la RADIOACTIVIDAD, usando el modelo: $Y_i = \beta_o + \beta_j + \epsilon_i$ si el pez i esta en el ACUARIO j; el parámetro β_j mide el efecto del ACUARIO j (j=1,..,3) sobre la RADIOACTIVIDAD. Escribe el criterio de los mínimos cuadrados en tres sumas que dependen de los tres acuarios. Usando la tabla 8.13 y tomando la media muestral $\bar{y}$ como estimador de β_o, deduzca el estimador de los mínimos cuadrados de los tres parámetros restantes.

b) Efectúe el test $H_o : \beta_3 = \beta_2 =$. Precise los supuestos que tuvo que hacer.

c) Sea un nuevo pez que toma los valores: LARGO: 180, LARGO SIN CABEZA: 152, ANCHO CABEZA: 40, ANCHO: 38, ANCHO HOCICO: 15, DIAMETRO OJOS: 12. Calcule C_1 y C_2 para este pez. Prediga su RADIOACTIVIDAD y dé un intervalo de confianza.
d) Si se supone que la variable RADIOACTIVIDAD $Y \sim Exp(\mu)$ y una distribución a priori $Exp(\theta)$ para μ ($\pi(\mu) = \theta exp(-\theta\mu)$ para μ positivo), dé la distribución a posteriori de μ dada la muestra de los 23 peces.
e) Tomando la función de perdida cuadratica, dé el estimador de Bayes.

	MEDIA	DESV. TIPICA	FACTOR .	FACTOR 2	FACTOR 3	FACTOR 4
VALORES PROPIOS			4.885	'493	'388	'128
% ACUMULADOS DE LOS VALORES PROPIOS			81.62	89.63	96.10	98.23
LARGO	.90.17	.7.99	0.947	0.226	0.099	0.126
LARGO	.90.17	.7.99	0.947	0.226	0.099	0.126
LARGO cIN CABEZA	.70.70	.5.69	0.939	0.264	0.128	-0.005
ANCHO CABEZA	42.78	4.80	0.959	0.133	0.121	0.045
ANCHO	39.30	4.57	0.922	-0.215	0.144	-0.283
ANCHO HOCICO	.3.57	2.54	0.816	0.071	-0.570	-0.040
DIAMETRO OJOS	9.74	0.96	0.817	-0.550	0.001	0.166

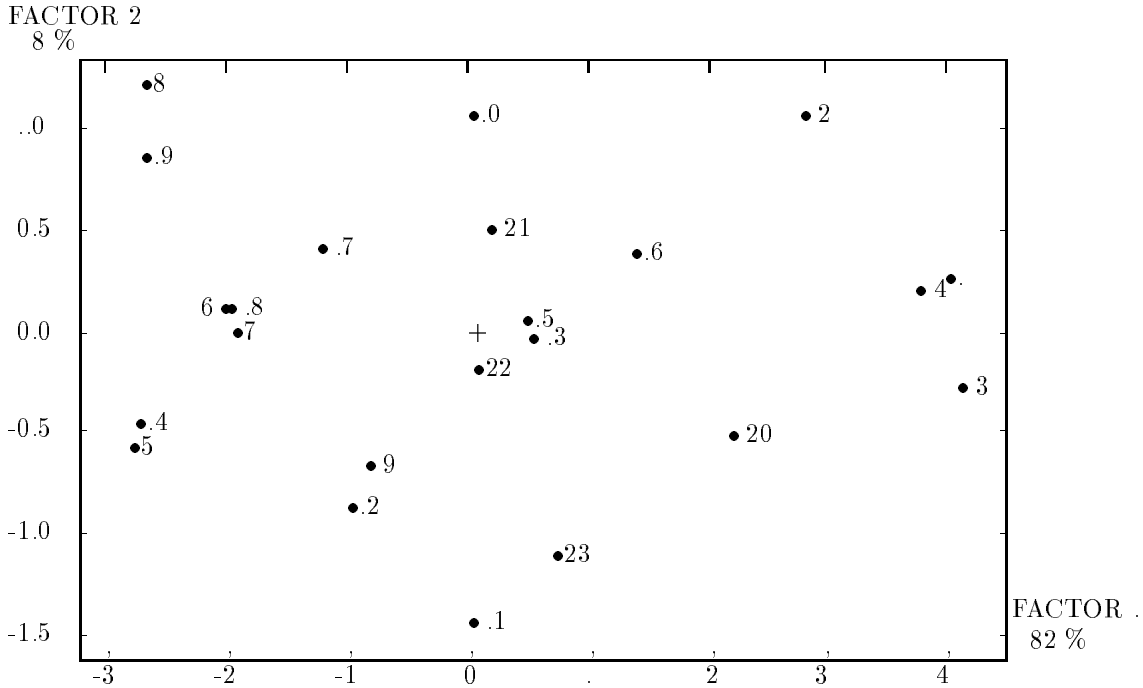
TABLA 8.11: Correlaciones de las variables sobre las 4 primeras componentes principales

VARIABLES	10	2	C_1	C_2
PESO	1.00	-.44	.98	.00
RADIOACTIVIDAD	-.44	1.00	-.41	.23
C_1	.98	-.41	1.00	.00
C_2	.00	.23	.00	1.00

TABLA 8.12: Matriz de correlaciones

ACUARIO	RADIOACTIVIDAD 1	2	3	TOTAL	PESO PESO
EFECTIVO	8	8	7	23	23
MEDIA	15.25	33.50	33.71	27.22	82.09
DESVIACION TIPICA	7.13	12.13	21.69	16.47	26.5

TABLA 8.13: Radiactividad


FACTOR 2
8 %
..0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
2
3
4
FACTOR .
82 %
8
.0
2
.9
21
.7
.6
4
6
.8
7
.5
.3
22
3
.4
5
20
9
.2
23
.1
GRAFICO 8.8