UNIVERSIDAD NACIONAL DE SAN AGUSTIN DE AREQUIPA

FACULTAD DE CIENCIAS NATURALES Y FORMALES

ESCUELA PROFESIONAL DE MATEMÁTICAS

ANALISIS DE DATOS MULTIVARIANTES MEDIANTE LA REDUCCIÓN DE LA DIMENSIONALIDAD

Tesis presentado por el Bachiller:
EDWIN GINO MORALES HUARACHI
Para optar el título profesional de:
Licenciado en Matemáticas.

AREQUIPA - 2015

INDICE

DEDICATORIA

Esta tesis se la dedico a Dios y a mi madre la Virgen de Chapi, quién supo guiarme por el buen camino, darme fuerzas para seguir adelante y no desmayar en los problemas que se presentaban, enseñándome a encarar las adversidades sin perder nunca la dignidad ni desfallecer en el intento.

A mi familia quienes por ellos soy lo que soy.

Para mis padres Ponciano y Basilia por su apoyo, consejos, comprensión, amor, ayuda en los momentos difíciles, y por brindarme los recursos necesarios para estudiar. Me han inculcado valores, principios, carácter, empeño, perseverancia, y coraje para conseguir mis objetivos.

A mis hermanos Aldo y Wanda por estar siempre presentes, acompañándome para lograr mis metas. A mi mejor amiga Sally Gómez quien ha sido y es mi motivación, inspiración y felicidad.

AGRADECIMIENTO

El presente trabajo de tesis primeramente me gustaría agradecerte a ti Dios por bendecirme para llegar hasta donde he llegado, porque hiciste realidad este sueño anhelado.

A mi asesor de tesis, Dr. Octavio Roque Roque por su esfuerzo y dedicación, quien con sus conocimientos, su experiencia, su paciencia y su motivación ha logrado en mí que pueda terminar mi tesis con éxito.

También me gustaría agradecer a mis profesores durante toda mi carrera profesional porque todos han aportado con un granito de arena a mi formación.

Son muchas las personas que han formado parte de mi vida profesional a las que me encantaría agradecerles su amistad, consejos, apoyo, ánimo y compañía en los momentos más difíciles de mi vida. Algunas están aquí conmigo y otras en mis recuerdos y en mi corazón, sin importar en donde estén quiero darles las gracias por formar parte de mí, por todo lo que me han brindado y por todas sus bendiciones.

Para ellos: Muchas gracias y que Dios los bendiga.

INTRODUCCIÓN

La información estadística proviene de respuestas o atributos, los cuales son observados o medidos sobre un conjunto de individuos u objetos denominados unidad de estudio, de modo que, cada respuesta o atributo está asociado con una unidad de estudio. Si tan sólo se registra un atributo por individuo, los datos resultantes son de tipo univariados, mientras que si más de una variable es registrada sobre una unidad de estudio, en esos casos, los datos tienen una estructura multivariada. No obstante, se pueden considerarse grupos de individuos de los cuales se obtienen muestras de datos multivariados para comparar algunas de sus características o parámetros. En una forma más general, los datos multivariados pueden proceder de varias poblaciones; donde el interés se dirige a la exploración descriptiva de las variables y la búsqueda de su interrelación dentro de los grupos y entre ellos. En otras palabras, los que suministran información sobre la interdependencia entre una o varias variables. Por lo que en este trabajo de tesis, se presenta el método de reducción de número de variables para obtener variables no dependientes, o sea, el objetivo general es obtener variables no correlacionadas a partir de las variables que suministran información sobre la interdependencia entre las variables, es decir, en una primera instancia, los datos multivariados siempre presentan correlación, el objetivo general es liberar esa correlación para simplificar el análisis de los mismos.

En el trabajo de recolección de la información sobre un campo determinado, uno de los problemas que enfrenta el investigador es la elección de las variables a medir. En un proceso de investigación, durante las etapas iniciales frecuentemente hay una escasa teoría sobre el campo a abordar; consecuentemente, el investigador recoge información sobre un número amplio de variables, que a su juicio son relevantes en el problema. En casos donde resultan muchas variables se presentan algunos problemas con la estimación de parámetros, así por ejemplo, con diez variables puede hacerse necesario 45 correlaciones, con 20 se pueden estimar 190 coeficientes de correlación, de tal forma que se hace necesario abocar alguna técnica que resuma la información contenida en las variables y facilite su análisis.

Por lo tanto, en el presente trabajo de tesis, como objetivo principal se tiene la reestructuración de un conjunto de datos multivariados mediante la reducción del número de variables. Esta es una metodología de tipo matemático para lo cual no es necesario asumir distribución probabilística alguna ni mucho menos la inferencia estadística. Lo que si es necesario el álgebra lineal de matrices.

El análisis descriptivo de datos multivariados mediante el método de reducción de variables tiene como objetivos, entre otros, los siguientes:

- Generar nuevas variables que expresen la información contenida en un conjunto de variables.
- Reducir la dimensión del espacio donde están inscritos los datos.
- Eliminar las variables (si es posible) que aporten poco al estudio del problema.
- Facilitar la interpretación de la información contenida en los datos.

El análisis por reducción de variables tiene como propósito central la determinación de pocos factores que retengan la mayor variabilidad contenida en los datos. Las nuevas variables poseen algunas características estadísticas deseables, tales como independencia y no correlación.

En el caso de no correlación entre las variables originales, la metodología de reducción de variables no tiene mucho que hacer, pues las componentes se corresponderían con cada variable por orden de magnitud en la varianza; es decir, la primera componente coincide con la primera variable de mayor varianza, la segunda componente con la variable de segunda mayor varianza, y así sucesivamente.

La estructuración del presente trabajo de tesis es como sigue: en el Capítulo I se inicia definiendo lo que es un vector aleatorio, vector de medias, matriz de covarianza, autovalores y autovectores de la matriz de varianza-covarianza y matriz de datos. También se estudia en este capítulo lo que es el tamaño de la muestra. El Capítulo II, es la parte medular el trabajo, lo cual se desarrolla desde dos puntos de vista, como es habitual en el desarrollo de estadística, es decir, se admite el enfoque intuitivo para conceptuar propiedades y característica sobre el tema a desarrollar, es así como se desarrollar la sección 2.2 desde el punto de vista intuitivo sobre componentes principales, luego se desarrolla el análisis de componentes principales para dos variables, y finalmente este concepto y propiedades de dos

variables se generaliza a n-variables, cuyo desarrollo se hace utilizando todo los elementos necesarios de argumento matemático.

Después de realizar el aspecto teórico sobre la reducción de variables, entramos al Capítulo III donde se estudia una aplicación a la industria manufacturera. Para lo cual se ha tomado datos proporcionados en Internet de INEI con sus respectivas fichas técnicas (anexo 1) y el instrumento aplicado para captar datos (anexo 2).

En todo el trabajo se usó como alternativa el software estadístico SPSS versión 22, para procesar los datos, ya que, este programa estadístico es más usual y bastante difundido entre los estadísticos y, al tiempo existen buena cantidad de libros o manuales sobre el paquete mencionado.

Edwin Gino Morales Huarachi

CAPÍTULO I

CONCEPTOS PRELIMINARES

1.1 INTRODUCCIÓN.

El concepto principal de este capítulo es el concepto de vector. Un conjunto de n datos numéricos de una variable puede representarse geométricamente asociando cada valor de la variable a una dimensión del espacio n dimensional, obteniendo un punto en ese espacio, con la única diferencia de que los valores que conforma al vector son variables aleatorias. A estas variables aleatorias se les aplicará los métodos estadísticos descriptivos y, para resumir información contenidas en ellas aplicaremos las operaciones concernientes a los vectores y matrices.

1.2 VECTOR ALEATORIO

En numerosas ocasiones estudiamos más de una variable asociado a un experimento aleatorio, como por ejemplo la velocidad de transmisión de un mensaje y la proporción de errores. De esta forma seremos capaces de estudiar no solo el comportamiento de cada variable por separado, sino las relaciones que pudieran existir en ellas. En otras palabras, el concepto de vector aleatorio nace como generalización natural de la noción de variable aleatoria unidimensional, al considerar simultáneamente el comportamiento aleatorio de varias características asociadas a un experimento.

Definición 1.1(vector aleatorio) Dado un espacio de probabilidades $(\Omega, \mathcal{A}, P)$ y el espacio probabilizable $(\mathcal{B}, R^n)$ con $\mathcal{B}$ σ-algebra de Borel, se dice que una aplicación

$\mathbf{X} = (x_1, x_2, \ldots, x_n)^t$

$$\mathbf{X} : \Omega \longrightarrow R^n$$

es un vector aleatorio si es medible, es decir $\mathbf{X}^{-1}(B) \in \mathcal{A}$ para todo $B \in \mathcal{B}$, por tanto cada uno de sus componentes x_i, $i = 1, \ldots, n$ es una variable aleatoria unidimensional.

Veamos algunos ejemplos de vectores aleatorios:

El vector (x_1, x_2) representa la temperatura máxima que puede alcanzar una resistencia y el tiempo que tarda en alcanzarla. (En el caso de bidimensional es más frecuente utilizar la notación (X, Y)).

Los componentes del vector

$$\mathbf{X} = (x_1, x_2, x_3, x_4, x_5, x_6)^t = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}_{6\times 1}$$

representa, respectivamente, la edad, peso, estatura, sexo, colesterol y triglicéridos de una persona. Esto es:

$$\mathbf{X} = \begin{pmatrix} edad \\ peso \\ estatura \\ sexo \\ colesterol \\ triglicéridos \end{pmatrix}$$

En teoría de probabilidades multidimensionales, un vector aleatorio se clasifica en dos tipos: vectores aleatorios discretos y vectores aleatorios continuos.

Un vector aleatorio es discreto, si todos sus componentes son variables aleatorios discretos.

Es continuo si todos sus componentes son variables aleatorias continuas.

Existe una propiedad muy importante, que $Var(\alpha v^t X) = \alpha^2 Var(v^t X)$ con $\alpha \in R$.

En efecto:

Por definición se tiene:

$$\begin{aligned} Var(\alpha v^t X) &= E[(\alpha v^t X) - E(\alpha v^t X)]^2 \\ &= E[\alpha(v^t X - E(v^t X))]^2 \\ &= E[\alpha^2 (v^t X - E(v^t X))^2] \\ &= \alpha^2 E[(v^t X - E(v^t X))^2] \\ &= \alpha^2 Var(v^t X) \text{ con } \alpha \in R. \end{aligned}$$

De modo que multiplicando por una constante podemos hacer la varianza tan grande o tan pequeña como queramos. Por ello es necesario normalizar, o sea, $\|v\| = 1$.

1.3 VECTOR DE MEDIAS

Definición 1.2 Dado un vector aleatorio $\mathbf{X} = (x_1, x_2, \ldots, x_n)^t$, se define su esperanza matemática como el vector

$$E[\mathbf{X}] = (E[x_1], E[x_2], \ldots, E[x_n])^t = \begin{pmatrix} E[x_1] \\ E[x_2] \\ \ldots \\ E[x_n] \end{pmatrix}.$$

A partir de la definición anterior son inmediatas las siguientes propiedades:

1. Si $\mathbf{X}_{n\times 1}$ es un vector aleatorio y $C_{n\times 1}$ es un vector de constantes, entonces $E[\mathbf{X} + C] = E[\mathbf{X}] + C$.
2. Si $\mathbf{X}_{n\times 1}$ e $\mathbf{Y}_{n\times 1}$ son vectores aleatorios, entonces $E[\mathbf{X} + \mathbf{Y}] = E[\mathbf{X}] + E[\mathbf{Y}]$.
3. Si $\mathbf{X}_{n\times 1}$ e $\mathbf{Y}_{n\times 1}$ son dos vectores aleatorios y $A_{p\times n}$ y $B_{p\times n}$ son dos matrices de constantes, entonces $E[A\mathbf{X} + B\mathbf{Y}] = AE[\mathbf{X}] + BE[\mathbf{Y}]$.
4. Si $\mathbf{X}_{n\times 1}$ es un vector aleatorio y $a_{n\times 1}$ es un vector de constantes, entonces $E[a^t\mathbf{X}] = a^t E[\mathbf{X}]$.

Ejemplo 1.1 Una moneda se lanza tres veces. Sea x_1 el número de caras conseguidos en los dos primeros lanzamiento. Sea y_2 el número de caras conseguidos en los dos últimos lanzamientos. Hallar la esperanza matemática $E[(X_1, X_2)]$.

Solución.

En primer lugar vamos a encontrar el espacio muestral asociado a este experimento aleatorio es:

$$\Omega = \{ccc, ccs, csc, scc, ssc, scs, css, sss\}$$

Los valores de las dos variables aleatorias son:

$$x_1 = 0, 1, 2;\ y_2 = 0, 1, 2$$

de donde:

$$R_X = \{0, 1, 2\} = R_Y$$

y,

$$R_X \times R_Y = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)\}$$

En la siguiente tabla mostramos la distribución de probabilidad correspondiente.

Tabla 1: Ley de probabilidad conjunta

X \ Y	0	1	2	Total
0	1/8	1/8	0	2/8
1	1/8	1/4	1/8	4/8
2	0	1/8	1/8	2/8
Total	2/8	4/8	2/8	1

Calculamos las esperanzas matemáticas univariantes o componentes.

$$\mu_X = E[X] = \sum_{X=0}^{2}(Xp(X,0) + Xp(X,1) + Xp(X,2)) = 0p(0,0) + 0p(0,1) + 0p(0,2)$$
$$+1p(1,0) + 1p(1,1) + 1p(1,2) + 2p(2,0) + 2p(2,1) + 2p(2,2)$$
$$= 0(1/8)+0(1/8)+0(0) + 1(1/8)+1(1/4)+1(1/8)+2(0)+2(1/8)+2(1/8)$$
$$= 8/8 = 1.$$

$$\mu_Y = E[Y] = \sum_{Y=0}^{2}(Yp(0,Y) + Yp(1,Y) + Yp(2,Y)) = 0p(0,0) + 0p(0,1) + 0p(0,2)$$
$$+1p(1,0) + 1p(1,1) + 1p(1,2) + 2p(2,0) + 2p(2,1) + 2p(2,2)$$
$$= 0(1/8)+0(1/8)+0(0) + 1(1/8)+1(1/4)+1(1/8)+2(0)+2(1/8)+2(1/8)$$
$$= 8/8 = 1.$$

Por tanto,

$$\mu = (E[X], E[Y]) = (1, 1).$$

1.4 MATRIZ DE COVARIANZA

La variabilidad de los datos y la información relativa a las relaciones lineales entre las variables se resumen en la matriz de covarianza.

Definición 1.3 Sea $A = (x_{ij})$, con $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$ una matriz aleatoria, esto es, todas sus componentes x_{ij} son variables aleatorias unidimensionales. Se define la esperanza matemática de A como matriz cuyas componentes son las correspondientes esperanzas de las variables x_{ij}. O sea $E[A] = (E[x_{ij}])$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$.

Teniendo en cuenta esta definición, las demostraciones de las siguientes propiedades son inmediatas.

1. Si $A_{n\times p}$ es una matriz aleatoria y $C_{n\times p}$ es una matriz de constantes, entonces $E[A + C] = E[A] + C$.
2. Si $A_{n\times p}$ y $B_{n\times p}$ son dos matrices aleatorias, entonces $E[A + B] = E[A] + E[B$.
3. Si $A_{n\times p}$ y $B_{m\times p}$ son dos matrices aleatorias y $X_{s\times n}$ y $Y_{s\times m}$ son dos matrices constantes, entonces $E[XA + YB] = XE[A] + YE[B]$.
4. Si $A_{n\times p}$ es una matriz aleatoria y si $X_{s\times n}$, $Y_{p\times l}$ y $Z_{s\times l}$ son tres matrices de constantes, entonces $E[XAY + Z] = XE[A]Y + Z$.

A partir de la definición anterior podemos introducir la definición de matriz de covarianza. En buena cuenta, es la generalización natural a dimensiones superiores del concepto de covarianza de una variable aleatoria bidimensional.

Definición 1.4 Sean $\mathbf{X_{p\times 1}}$ e $\mathbf{Y_{q\times 1}}$ dos vectores aleatorios. Se define la covarianza entre ambos vectores, que denotaremos por Cov[X , Y], como la esperanza matemática de la matriz aleatoria $(X-E[X])(Y - E[Y])^t$.

La definición anterior en forma desarrollada se tiene como:

$$Cov[X , Y] = E[(X-E[X])(Y - E[Y])^t]$$

$$= \mathrm{E}\left[\begin{pmatrix} X_1 - E[X_1] \\ \dots \\ X_i - E[X_i] \\ \dots \\ X_p - E[X_p] \end{pmatrix} \left(Y_1 - E[Y_1], \dots, Y_j - E\left[Y_j\right], \dots, Y_q - E[Y_q]\right)\right]$$

$$= \mathrm{E}\begin{bmatrix} (X_1 - E[X_1])(Y_1 - E[Y_1]) & \dots & \left((X_1 - E[X_1])(Y_j - E[Y_j])\right) & \dots & \left((X_1 - E[X_1])(Y_q - E[Y_q])\right) \\ \dots & \dots & & \dots & \dots \\ (X_i - E[X_i])(Y_1 - E[Y_1]) & \dots & (X_i - E[X_i])(Y_j - E[Y_j]) & \dots & (X_i - E[X_i])(Y_q - E[Y_q]) \\ \dots & \dots & \dots & \dots & \dots \\ (X_p - E[X_p])(Y_1 - E[Y_1]) & \dots & (X_p - E[E_p])(Y_q - E[Y_q]) & \dots & (X_p - E[X_p])(Y_q - E[Y_q]) \end{bmatrix}$$

$$= \begin{pmatrix} Cov[X_1, Y_1] & \dots & Cov[X_1, Y_j] & \dots & Cov[X_1, Y_q] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_i, Y_1] & \dots & Cov[X_i, Y_j] & \dots & Cov[X_i, Y_q] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_p, Y_1] & \dots & Cov[X_i, Y_j] & \dots & Cov[X_p, Y_q] \end{pmatrix}$$

Veamos algunas propiedades y consecuencias inmediatas de la definición anterior.

1. Si $X_{p\times1}$ e $Y_{q\times1}$ son dos vectores aleatorios, entonces $Cov[X, Y] = E[XY^t] - (E[X])(E[Y])^t$.
2. Si $X_{p\times1}$ e $Y_{q\times1}$ son dos vectores aleatorios y $a_{p\times1}$ y $b_{q\times1}$ son dos vectores de constantes, entonces $Cov[X- a, Y - b] = Cov[X, Y]$.
3. Si $X_{p\times1}$ e $Y_{q\times1}$ son dos vectores aleatorios y $A_{s\times p}$ y $B_{t\times q}$ son dos matrices de constantes, entonces $Cov[AX, BY] = ACov[X, Y]B^t$.
4. Si $X_{p\times1}$ e $Y_{q\times1}$ son dos vectores aleatorios y si $a_{p\times1}$ y $b_{q\times1}$ son dos vectores de constantes, entonces $Cov[a^tX, b^tY] = a^tCov[X, Y]b$.

En el caso particular cuando X = Y tenemos la matriz de varianza-covarianza del vector aleatorio **X**. Esta matriz es cuadrada y simétrica de orden n, donde los términos diagonales son la varianza y los no diagonales son las covarianzas. Una matriz de varianza-covarianza se denotará por Σ.

Definición 1.5 Sea $\mathbf{X_{p\times1}}$ un vector aleatorio. Una matriz de varianzas-covarianzas de **X**, dado por $Cov[X] = \Sigma$, se define como la esperanza matemática de la matriz aleatoria $(X-E[X])(X – E[X])^t$. Es decir:

$$Cov[X] = E[(X-E[X])(X – E[X])^t] = \Sigma.$$

La definición anterior en forma desarrollada es como sigue:

$$\text{Cov}[X] = E[(X-E[X])(X - E[X])^t]$$

$$= E\left[\begin{pmatrix} X_1 - E[X_1] \\ \dots \\ X_i - E[X_i] \\ \dots \\ X_p - E[X_p] \end{pmatrix}\left(X_1 - E[X_1], \dots, X_i - E[X_i], \dots, X_p - E[X_p]\right)\right]$$

$$= E\begin{bmatrix} (X_1 - E[X_1])^2 & \dots & \left((X_1 - E[X_1])(X_i - E[X_i])\right) & \dots & \left((X_1 - E[X_1])(X_p - E[X_p])\right) \\ \dots & \dots & \dots & \dots & \dots \\ (X_i - E[X_i])(X_1 - E[X_1]) & \dots & (X_i - E[X_i])^2 & \dots & (X_i - E[X_i])(X_p - E[X_p]) \\ \dots & \dots & \dots & \dots & \dots \\ (X_p - E[X_p])(X_1 - E[X_1]) & \dots & (X_p - E[E_p])(X_i - E[X_i]) & \dots & (X_p - E[X_p])^2 \end{bmatrix}$$

$$= \begin{pmatrix} Var[X_1] & \dots & Cov[X_1, X_i] & \dots & Cov[X_1, X_p] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_i, X_1] & \dots & Var[X_i] & \dots & Cov[X_i, X_p] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_p, X_1] & \dots & Cov[X_p, X_i] & \dots & Var[X_p] \end{pmatrix}$$

Las siguientes propiedades son consecuencia inmediatas de la definición anterior y de lo expuesto para el operador covarianza anteriormente introducido.

1. Si $\mathbf{X_{p\times 1}}$ es un vector aleatorio, entonces $\text{Cov}[X] = E[XX^t] - (E[X])(E[X])^t$

2. Si $\mathbf{X_{p\times 1}}$ es un vector aleatorio y si $a_{p\times 1}$ es un vector de constantes, entonces

   $\text{Cov}[X-a] = \text{Cov}[X]$.

3. Si $\mathbf{X_{p\times 1}}$ es un vector aleatorio y $A_{s\times p}$ es un matriz de constantes, entonces

   $\text{Cov}[A\mathbf{X}] = A\text{Cov}[\mathbf{X}]A^t$

4. Si $\mathbf{X_{p\times 1}}$ es un vector aleatorio y $a_{p\times 1}$ es un vector de constantes, entonces

   $\text{Var}[a^t\mathbf{X}] = a^t\text{Cov}[\mathbf{X}]a$.

Una consecuencia inmediata de la matriz de varianza-covarianza es la matriz de correlaciones. Así:

$$\Sigma = \begin{pmatrix} Var[X_1] & \dots & Cov[X_1, X_i] & \dots & Cov[X_1, X_p] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_i, X_1] & \dots & Var[X_i] & \dots & Cov[X_i, X_p] \\ \dots & \dots & \dots & \dots & \dots \\ Cov[X_p, X_1] & \dots & Cov[X_p, X_i] & \dots & Var[X_p] \end{pmatrix}$$

Entonces, la matriz de correlación es.

$$R = \begin{pmatrix} 1 & \cdots & \frac{Cov[X_1,X_i]}{\sqrt{Var[X_1]}\sqrt{Var[X_i]}} & \cdots & \frac{Cov[X_1,X_p]}{\sqrt{Var[X_1]}\sqrt{Var[X_p]}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{Cov[X_i,X_1]}{\sqrt{Var[X_i]}\sqrt{Var[X_1]}} & \cdots & 1 & \cdots & \frac{Cov[X_i,X_p]}{\sqrt{Var[X_i]}\sqrt{Var[X_p]}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{Cov[X_p,X_1]}{\sqrt{Var[X_p]}\sqrt{Var[X_1]}} & \cdots & \frac{Cov[X_p,X_i]}{\sqrt{Var[X_p]}\sqrt{Var[X_i]}} & \cdots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \cdots & \rho_{1i} & \cdots & \rho_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{i1} & \cdots & 1 & \cdots & \rho_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{pi} & \cdots & \rho_{pi} & \cdots & 1 \end{pmatrix}$$

Para todo i se cumple $\rho_{1i} = \rho_{i1}$. En general se tiene:

$$\rho_{ij} = \rho_{ji}; \text{ para todo i, j = 1, 2, …, n}$$

y,

$$-1 \leq \rho_{ij} \leq 1. \text{ Para todo i, j.}$$

Ejemplo 1.2 En el estado de Ceuta y Melilla de España, existen empresas de un sector de actividades que conforman grupos estratégicos que aprovechando la política de liberalización, uso intensivo de la tecnología y la información y la globalización de los mercados financieros, según sus sectores productivas se tiene la siguientes informaciones de sus movimientos mercantilistas y las ventajas hacia el usuario, según el anuario 1994 del País.

Tabla 2: Ventas y sectores productivos

N°	EMPRESA	VENTAS (Miles de Euros)	BENEFICIOS (Miles de Euros)
1	El Corte Ingles	775.104	23.795
2	Iberdrola	775.218	58.778
3	Repsol Comercial	700.963	1.531
4	Seat	674.063	-12.756
5	Tabacalera	631.003	14.729

6	FASA Renault	537.744	9.059
7	Repsol Petroleo	489.155	12.541
8	Pryca	448.465	13.495
9	Iberia	445.853	-34.824

Calcular la matriz de covarianza y correlación.

Solución.

Sabemos que la covarianza y la correlación son:

$$\text{cov(Ventas, Beneficios)} = \frac{\sum(ventas)(beneficios)}{9} - \overline{ventas} * \overline{beneficios}$$

$\overline{ventas}$ = promedio de ventas

$\overline{beneficios}$ = promedio de beneficios

y,

Correlación (Ventas, Beneficios) = r =

$$= \frac{9(\sum ventas)(beneficios) - (\sum ventas)(\sum beneficios)}{\sqrt{[9(\sum ventas^2) - (\sum ventas)^2][9(\sum beneficios^2) - (\sum beneficios)^2]}}$$

O sea,

$$\text{cov(Ventas, Beneficios)} = \frac{67309.80}{9} - (608.62)(9.59)$$

$$= 7478.87 - 5836.67$$

$$= 1642.2$$

$$\text{var(Ventas)} = \frac{3473970.98}{9} - (608.62)^2$$

$$= 385996.78 - 370418.30$$

$$= 17527.61$$

desviación estándar = 132.81

var(Beneficios) = 651.09

desviación estándar = 25.52

Luego la matriz de covarianza es:

$$\Sigma = \begin{pmatrix} 132.39 & 1642.2 \\ 1642.2 & 25.52 \end{pmatrix}$$

y la matriz de coeficiente de correlación es:

$$R = \begin{pmatrix} 1 & 0.4860 \\ 0.4860 & 1 \end{pmatrix}.$$

1.4.1 AUTOVALORES Y AUTOVECTORES DE LA MATRIZ DE VARIANZA-COVARIANZA "Σ".

Dado una matriz $\Sigma \in K^{n\times n}$, e I una matriz identidad de $K^{n\times n}$. Entonces

$$\det(\Sigma - \lambda I) = 0 = |\Sigma - \lambda I|$$

se llama **polinomio característico** de Σ.

Proposición 1.1 Sea $\lambda \in K$ un autovalor de la matriz $\Sigma \in K^{n\times n}$ si y sólo si, λ es raíz del polinomio característico de la matriz Σ.

Demostración.

Si λ es un autovalor de Σ si y solo si $(\Sigma - \lambda I)v = \mathbf{0}$ para algún $v \neq 0$, si y solo si

$$|\Sigma - \lambda I| = 0.$$

Es decir, λ es un autovalor de Σ si y solo λ es raíz de un polinomio característico.

Ejemplo 1.3 Sea

$$\Sigma = \begin{bmatrix} -11 & -10 & 5 \\ 0 & 4 & 0 \\ -15 & -10 & 9 \end{bmatrix} \in R^{3\times 3}$$

una matriz de varianza-covarianza. Entonces el polinomio característico de Σ es

$$\begin{vmatrix} -11-\lambda & -10 & 5 \\ 0 & 4-\lambda & 0 \\ -15 & -10 & 9-\lambda \end{vmatrix} = 0,$$

si y solo si

$$(4 - \lambda)(\lambda^2 + 2\lambda - 24) = 0 \Leftrightarrow -(\lambda - 4)^2(\lambda + 6) = 0$$

$$\Leftrightarrow \lambda_1 = -6 \text{ y } \lambda_2 = 4.$$

Estas raíces del polinomio característico son autovalores de Σ.

Las raíces $\lambda_1 = -6$ y $\lambda_2 = 4$ cuyas multiplicidades algebraicas son 1 y 2 respectivamente, entonces se pueden calcular los correspondientes autovectores de estos autovalores. Así:

- Para λ_1= - 6 tenemos:

$$\Sigma\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}-11 & -10 & 5\\ 0 & 4 & 0\\ -15 & -10 & 9\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} = -6\begin{bmatrix}x\\y\\z\end{bmatrix} \Leftrightarrow \begin{bmatrix}-11x-10y+5z\\ 4y\\ -15x-10y+9z\end{bmatrix} = -6\begin{bmatrix}x\\y\\z\end{bmatrix}$$

$$\Leftrightarrow \begin{cases}-11x-10y+5z = -6x\\ 4y = -6y\\ -15x-10y+9z = -6z\end{cases}$$

$$\Leftrightarrow \begin{cases}-5x-10y+5z=0\\ 10y=0\\ -15x-10y+15z=0\end{cases}$$

$$\Leftrightarrow \begin{cases}-x+z=0\\ y=0\\ -x+z=0\end{cases}$$

$$\Leftrightarrow \begin{cases}z=x\\ y=0\end{cases}$$

Luego el primer autovector es:

$$V_1 = \begin{bmatrix}x\\0\\x\end{bmatrix}, \text{ para todo real } x \neq 0.$$

- Para λ_2 = 4 tendremos:

$$\Sigma\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}-11 & -10 & 5\\ 0 & 4 & 0\\ -15 & -10 & 9\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} = 4\begin{bmatrix}x\\y\\z\end{bmatrix} \Leftrightarrow \begin{bmatrix}-11x-10y+5z\\ 4y\\ -15x-10y+9z\end{bmatrix} = 4\begin{bmatrix}x\\y\\z\end{bmatrix}$$

$$\Leftrightarrow \begin{cases}-11x-10y+5z = 4x\\ 4y = 4y\\ -15x-10y+9z = 4z\end{cases}$$

$$\Leftrightarrow \begin{cases}-15x-10y+5z=0\\ 0=0\\ -15x-10y+5z=0\end{cases}$$

$$\Leftrightarrow \begin{cases}-3x-2y+z=0\\ 0=0\\ -3x-2y+z=0\end{cases}$$

$$\Leftrightarrow z = 3x + 2y$$

Luego el segundo autovector es:

$$V_1 = \begin{bmatrix} x \\ y \\ 3x + 2y \end{bmatrix}, \text{ para todo real x , y} \neq 0.$$

1.5 MATRIZ DE DATOS

La matriz de datos es un modo de ordenar los datos de manera natural que sea particularmente visible la estructura tripartita de los datos.

Los datos se arreglan de tal forma que las unidades de estudio (U = 1, 2, 3, …, L) se ubican en los renglones y cada variable (V = 1, 2, 3, …,K) en las columnas.

1) Si se desea conocer todas las características de una unidad específica se recorre todo el renglón.

2) Si se desea conocer como se distribuyen las unidades en los distintos valores de una variable, se recorre la columna.

Las celdas están formadas por las intersecciones de los renglones y las columnas contienen los valores (r).

1) Cada valor (r) es la respuesta de la i-ésima unidad en la k-ésima variable. A la inversa: toda combinación (U_i , V_k) define en la matriz un punto (r_{ik}).

2) La falta de valor (de un valor de los predeterminados) en una celda es denominado “sin datos”. En consecuencia, la ausencia de valor se representa mediante un código (99, 999, etc). Es una situación frecuente en las matrices de datos. Una forma de valorar una matriz es por la cantidad de “sin datos” que tiene.

Ejemplo 1.4 Supongamos que tenemos las medidas corporales de 33 solicitantes para ingresar a la policía nacional. Los datos se tomaron de Métodos Multivariados de Dallas E. Johnson 1998-pág 10. Las variables son medidas en centímetros, fueron la estatura (EST), la estatura sentada (ESTSEN), la longitud del brazo (BRAZO), la longitud del antebrazo (ANTEB), el ancho de la mano (MANO), la longitud del muslo (MUSLO), la longitud de la parte inferior de la pierna (PIERNA) y la longitud del pie (PIE). A partir de estos datos, se crearon dos variables adicionales: la razón de la longitud del antebrazo a la del brazo multiplicada por 100 (BRACH) y la razón de la parte inferior de la pierna a la del muslo multiplicado por 100 (TIBIO). Por tanto, BRACH = 100. ANTEB/BRAZO, y TIBIO= 100.PIERNA/MUSLO:

Tabla 3: Medidas del cuerpo en solicitantes al departamento de policía

ID	EST	ESTSEN	BRAZO	ANTEB	MANO	MUSLO	PIERNA	PIE	BRACH	TIBIO
1	165.4	88.7	31.8	28.1	18.7	40.3	38.9	6.7	88.36	96.53
2	169.8	90.0	32.4	29.1	18.3	43.3	42.7	6.4	89.81	98.61
3	170.7	87.7	33.6	29.5	20.7	43.7	41.1	7.2	87.80	84.05
4	170.9	87.1	31.0	28.2	18.6	43.7	40.6	6.7	90.97	92.91
5	157.5	81.3	32.1	27.3	17.5	38.1	39.6	6.6	85.05	103.94
6	165.9	88.2	31.8	29.0	18.6	42.0	40.6	6.5	91.19	96.67
7	158.7	86.1	30.6	27.8	18.4	40.0	37.0	5.9	90.85	92.50
8	166.0	88.7	30.2	26.9	17.5	41.6	39.0	5.9	89.07	93.75
9	158.7	83.7	31.1	27.1	18.3	38.9	37.5	6.1	87.14	96.40
10	161.5	81.2	32.3	27.8	19.1	42.8	40.1	6.2	86.07	93.69
11	167.3	88.6	34.8	27.3	18.3	43.1	41.8	7.3	78.45	96.98
12	167.4	83.2	34.3	30.1	19.2	43.4	42.2	6.8	87.76	97.24
13	159.2	81.5	31.0	27.3	17.5	39.8	39.6	4.9	88.06	99.50
14	170.0	87.9	34.2	30.9	30.9	43.1	43.7	6.3	90.35	101.39
15	166.3	88.3	30.6	28.8	28.8	41.8	41.0	5.9	94.12	98.09
16	169.0	85.6	32.6	28.8	28.8	42.7	42.0	6.0	88.34	98.36
17	156.2	81.6	31.0	25.6	25.6	44.2	39.0	5.1	82.58	88.24

18	159.6	86.6	32.7	25.4	25.4	42.0	37.5	5.0	77.68	89.29
19	155.0	82.0	30.3	26.6	26.6	37.9	36.1	5.2	87.79	95.25
20	161.1	84.1	29.5	26.6	26.6	38.6	38.2	5.9	90.17	98.96
21	170.3	88.1	34.0	29.3	29.3	43.2	41.4	5.9	86.18	95.83
22	167.8	83.9	32.5	28.6	28.6	43.3	42.9	7.2	88.00	99.08
23	163.1	88.1	31.7	26.9	26.9	40.1	39.0	5.9	84.86	97.26
24	165.8	87.0	33.2	26.3	26.3	43.2	40.7	5.9	79.22	94.21
25	175.4	89.6	35.2	30.1	30.1	45.1	44.5	6.3	85.51	98.67
26	159.8	85.6	31.5	27.1	27.1	42.3	39.0	5.7	86.03	92.20
27	166.0	84.9	30.5	28.1	28.1	41.2	43.0	6.1	92.13	104.37
28	161.2	84.1	32.8	29.2	29.2	42.6	41.1	5.9	89.02	96.48
29	160.4	84.3	30.5	27.8	27.8	41.0	39.8	6.0	91.15	97.07
30	164.3	85.0	35.0	27.8	27.8	47.2	42.4	5.0	79.43	89.83
31	165.5	83.4	36.2	28.6	28.6	45.0	42.3	5.6	79.01	94.00
32	167.2	84.3	33.6	27.1	27.1	46.0	41.6	5.6	80.65	90.43
33	167.2	89.6	33.5	29.7	29.7	45.2	44.0	5.2	88.66	97.35

1.6 DETERMINACIÓN DEL TAMAÑO ADECUADO DE UNA MUESTRA.

Para tener una aplicación coherente y explicativa, es necesario tener un adecuado tamaño de muestra representado por “n”, ya que se conoce el número total de encuestados. Una fórmula

muy conocida y extendida que orienta sobre el cálculo del tamaño de la muestra para datos globales es la siguiente:

$$n = \frac{k^2 * N * p * q}{e^2 * (N-1) + k^2 * p * q}$$

Donde:

N: es el tamaño de la población o universo (número total de posibles encuestados).

k: es una constante que depende del nivel de confianza que asignemos. Los valores de k se obtienen de la tabla de la distribución normal estándar N(0,1). (Por tanto si pretendemos obtener un nivel de confianza del 95% necesitamos poner en la fórmula k=1,96)

e: es el error muestral deseado, en tanto por uno. El error muestral es la diferencia que puede haber entre el resultado que obtenemos preguntando a una muestra de la población y el que obtendríamos si preguntáramos al total de ella. Por ejemplo: si los resultados de una encuesta dicen que 100 personas comprarían un producto y tenemos un error muestral del 5% comprarán entre 95 y 105 personas.

p: proporción de individuos que poseen en la población la característica de estudio. Este dato es generalmente desconocido y se suele suponer que p = q = 0.5 que es la opción más segura.

q: proporción de individuos que no poseen esa característica, es decir, es 1-p.

n: tamaño de la muestra adecuada.

Altos niveles de confianza y bajo margen de error no significan que la encuesta sea de mayor confianza o esté más libre de error necesariamente; antes es preciso minimizar la principal fuente de error que tiene lugar en la recogida de datos.

En nuestro caso se tiene:

N = 1220 información encontrada por IN EI.

k = 1.96 (5%)

p = q = 0.5

e = 0.05.

Reemplazando estos valores en la fórmula anterior se tiene:

$$n = \frac{(1.96)^2 * 1220 * 0.5 * 0.5}{(0.05)^2 * 1219 + (1.96)^2 * 0.5 * 0.5}$$

$$= \frac{1171.69}{3.0475 + 0.9604}$$

$$= 292.34 \cong 293.$$

CAPÍTULO II

ANALISIS DE REDUCCIÓN DE VARIABLES

2.1 INTRODUCCIÓN

El análisis de reducción de variables es justamente el estudio de componentes principales, que tiene como objetivo transformar un conjunto de variables, a las que denominaremos variables originales, en un nuevo conjunto de variables denominadas componentes principales. Esta últimas se caracterizan por estar incorrelacionadas entre si.

El enfoque que se presentará en este trabajo de tesis sobre análisis de reducción de variables, en primera instancia un enfoque intuitivo, luego se hará un estudio analítico o la formalización correspondiente.

2.2 CONCEPTO GENERAL DE COMPONENTES PRINCIPALES

En muchas ocasiones el investigador se enfrenta a situaciones en las que, para analizar un fenómeno, dispone de información de muchas variables que están correlacionadas entre sí en mayor o menor grado. Estas correlaciones son como un velo que impide evaluar adecuadamente el papel que juega cada variable en el fenómeno estudiado. Por lo que, el método de reducción de variables permite pasar a un nuevo conjunto de variables llamadas componentes principales, que se caracterizan por no ser correlacionadas, o sea, variables incorrelacionadas entre sí y que pueden ordenarse de acuerdo con la información que llevan incorporada.

Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada en dicha componente. Por esta razón se selecciona como primera componente aquella que tenga mayor varianza, mientras que, por el contrario, la última es la de menor varianza. En general, la extracción de componentes principales se efectúa sobre variables tipificadas para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en desviaciones respecto a la media.

Si p variables están tipificadas, la suma de las varianzas es igual a p, ya que la varianza de una variable tipificada es por definición igual a 1.

El nuevo conjunto de variables que se obtiene por el método de componentes principales es igual en números al de variables originales. Es importante destacar que la suma de sus varianzas es igual a la suma de las varianzas de las variables originales.

Las diferencias entre ambos conjuntos de variables estriban en que, como ya se ha indicado, las componentes principales se calculan de forma que estén incorrelacionadas entre sí. Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes.

Si las variables originales estuvieron completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

Si la correlación muestral es nula entre el conjunto de variables, entonces las componentes principales coincidirán exactamente con las variables originales. Así pues, para aplicar este análisis hay que partir del supuesto de que las variables están correlacionadas entre sí

Desde el punto de vista de su aplicación, el método de componentes principales es considerado como un método de reducción, es decir, un método que permite reducir la dimensión del número de variables que inicialmente se considerado en el análisis.

Es importante destacar que las componentes principales se expresan como una combinación lineal de las variables originales.

2.3 ANALISIS DE COMPONENTES PRINCIPALES DE DOS VARIABLES

Para comprender mejor lo dicho en la sección anterior, ilustraremos para el caso bidimensional mediante un ejemplo, ya que, la formalización analítica se hará para el caso general de n- variables.

Para mayor facilidad en cuanto a la operacionalidad matemática y presentación gráfica y, a modo de familiarizarse con el programa estadístico SPSS, versión 22, utilizaremos en este ejemplo el mencionado programa.

Ejemplo 2.1 Consideremos un caso de dos variables para dar una visión intuitiva del método. Para tal efecto se tiene la tabla siguiente:

Table 4: Empresas y Ventas-Beneficios

N°	**EMPRESA**	**VENTAS (Miles de Euros)**	**BENEFICIOS (Miles de Euros)**

1	El Corte Ingles	775.104	23.795
2	Iberdrola	775.218	58.778
3	Repsol Comercial	700.963	1.531
4	Seat	674.063	-12.756
5	Tabacalera	631.003	14.729
6	FASA Renault	537.744	9.059
7	Repsol Petroleo	489.155	12.541
8	Pryca	448.465	13.495
9	Iberia	445.853	-34.824

En primer lugar, veamos el diagrama de puntos mediante el programa SPSS.

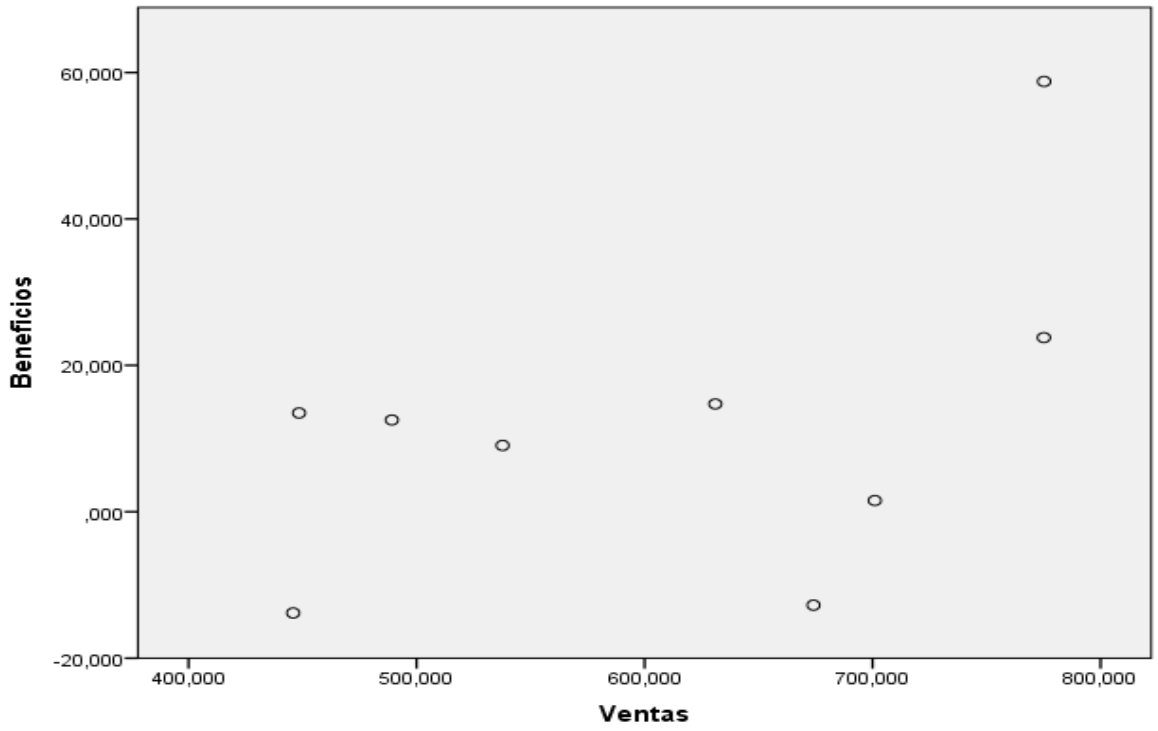

Fig 1: Ilustración gráfica de Ventas-beneficios

Tabla 5: Correlaciones de ventas-beneficios

		Ventas	Beneficios
Ventas	Correlación de Pearson	1	,495
	Sig. (bilateral)		,175

	N	9	9
Beneficios	Correlación de Pearson	,495	1
	Sig. (bilateral)	,175	
	N	9	9

Observamos que la correlación positiva es igual a 0.495. Esto significa que existe una correlación moderada entre las variables de beneficios y ventas en sentido positivo. Podemos tipificar las dos variables, para lo cual necesitamos conocer la media y desviación estándar y así:

$$z = \frac{Ventas - \overline{Ventas}}{\sigma_{Ventas}}, \quad z = \frac{Beneficios - \overline{Beneficios}}{\sigma_{Beneficios}}$$

Tabal 6: Estadísticos Ventas-beneficios

		Ventas	Beneficios
N	Válido	9	9
	Perdidos	0	0
Media		608,61867	11,92756
Desviación estándar		132,391864	21,607924
Varianza		17527,606	466,902

Luego se tiene los valores tipificados o estandarizados.

Tabla 7: Valores tipificado de ventas-beneficios.

N°	EMPRESA	VENTAS (Miles de Euros) Tipificada	BENEFICIOS (Miles de Euros) Tipificada
1	El Corte Ingles	1.257	0.556
2	Iberdrola	1.258	1.927
3	Repsol Comercial	0.697	-0.316
4	Seat	0.494	-0.875
5	Tabacalera	0.169	0.201
6	FASA Renault	-0.535	-0.020
7	Repsol Petroleo	-0.902	0.115
8	Pryca	-1.209	0.152
9	Iberia	-1.229	-1.740

El diagrama de dispersión de los valores tipificados se observa en el siguiente gráfico, ubicados o distribuidos entre los campos negativos y positivos.

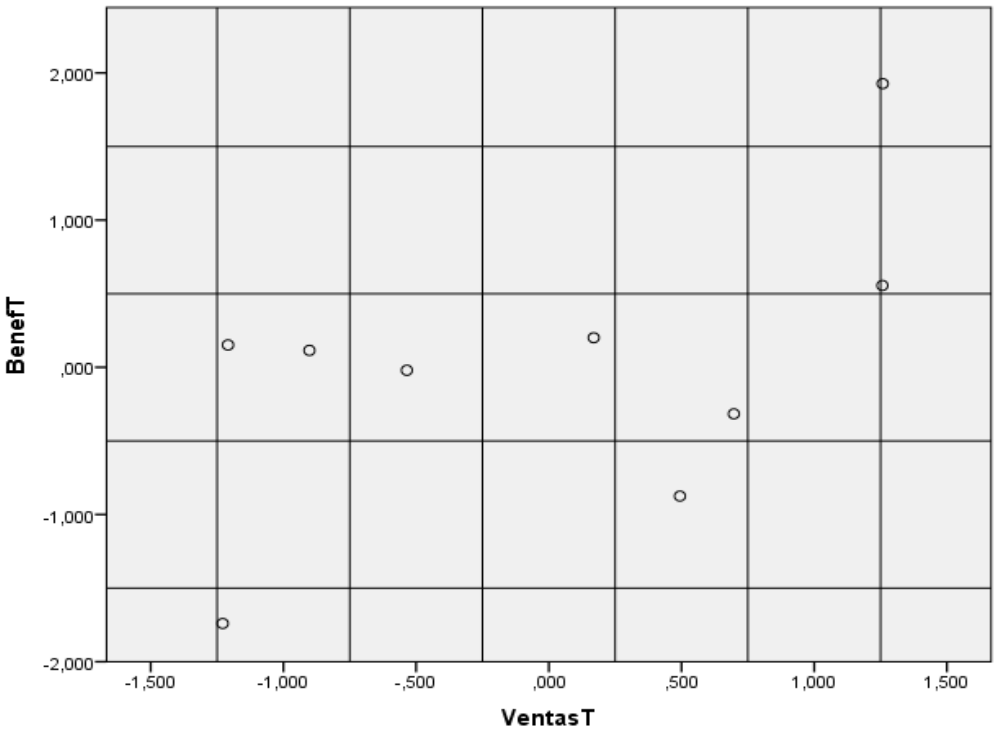

Fig 2: Ilustración gráfica de Ventas-beneficios tipificados

Cuando se tipifican las observaciones, entonces la matriz de covarianzas es precisamente la matriz de correlación y, por lo tanto, la varianza de cada variable tipificada es igual a 1.

Tabla 8: Correlaciones de los valores tipificados.

		VentasT	BenefT
VentasT	Correlación de Pearson	1	,546
	Sig. (bilateral)		,128
	N	9	9
BenefT	Correlación de Pearson	,546	1
	Sig. (bilateral)	,128	
	N	9	9

Y, la matriz de correlación para las dos variables tipificadas es:

$$R = \begin{bmatrix} 1 & 0.546 \\ 0.546 & 1 \end{bmatrix}$$

Al aplicar el método de componentes principales, la suma de las varianzas de todas las componentes principales es igual a la suma de las varianzas de las variables originales. En consecuencia, en el caso de dos variables tipificadas esta suma debe ser igual a 2.

Tabla 9: Correlaciones de Pearson

		Ventas	Beneficios
Ventas	Correlación de Pearson	1	,495
	Sig. (bilateral)		,175
	Suma de cuadrados y productos vectoriales	140220,844	11338,716
	Covarianza	17527,606	1417,339
	N	9	9
Beneficios	Correlación de Pearson	,495	1
	Sig. (bilateral)	,175	
	Suma de cuadrados y productos vectoriales	11338,716	3735,219
	Covarianza	1417,339	466,902
	N	9	9

La matriz de covarianza cuando las variables no son tipificadas es:

$$\Sigma = \begin{bmatrix} 17527.61 & 1417.34 \\ 1417.34 & 466.90 \end{bmatrix}$$

La aplicación del procedimiento de componentes principales requiere calcular los autovalores y los autovectores de la matriz de covarianza. Para tal efecto se tiene:

$$\begin{vmatrix} 17527.61 - \lambda & 1417.34 \\ 1417.34 & 466.90 - \lambda \end{vmatrix} = 0 \Leftrightarrow (17527.61-\lambda)(466.90-\lambda) - (1417.34)^2 = 0$$

$$\Leftrightarrow 8183641.11 - 17994.51\lambda + \lambda^2 - 2008852.68 = 0$$

$$\Leftrightarrow \lambda^2 - 17574.51\lambda + 6174788.43 = 0$$

$$\Leftrightarrow \begin{cases} \lambda_1 = \frac{17574.51+\sqrt{308863401.7-24699153.72}}{2(1)} \\ \lambda_2 = \frac{17574.51-\sqrt{308863401.7-24699153.72}}{2(1)} \end{cases}$$

$$\Leftrightarrow \begin{cases} \lambda_1 = 17215.84 \\ \lambda_2 = 358.670 \end{cases}$$

Son los autovalores de la matriz Σ.

Cuando se tipifican las observaciones, entonces la matriz de covarianza es precisamente la matriz de correlaciones y, por lo tanto, la varianza de cada variable tipificada es igual a 1. En este caso tendremos la matriz

$$R = \begin{bmatrix} 1 & 0.546 \\ 0.546 & 1 \end{bmatrix}$$

luego el polinomio característico es:

$$\begin{vmatrix} 1 - \lambda & 0.546 \\ 0.546 & 1 - \lambda \end{vmatrix} = 0 \Leftrightarrow (1-\lambda)^2 - 0.2981 = 0$$

$$\Leftrightarrow 1-2\lambda + \lambda^2 - 0.2981 = 0$$

$$\Leftrightarrow \lambda^2 - 2\lambda + 0.7019 = 0$$

$$\Leftrightarrow \begin{cases} \lambda_1 = \frac{2+\sqrt{4-2.81}}{2(1)} \\ \lambda_2 = \frac{2-\sqrt{4-2.81}}{2(1)} \end{cases}$$

$$\Leftrightarrow \begin{cases} \lambda_1 = 1.545 \\ \lambda_2 = 0.455 \end{cases}$$

Y, según el programa SPSS tenemos.

Tabla 10: Autovalores iniciales mediante SPSS

	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	**1,546**	77,313	77,313	1,546	77,313	77,313
2	**,454**	22,687	100,000			

La varianza de cada componente principal es igual al valor cada autovalor asociado. En el caso de bidimensional tipificada, la varianza de la primera componente principal es igual a 1 (que es la varianza de una de las variables tipificadas), más el 0.54603, que es el coeficiente entre las dos variables. La segunda componente principal su varianza es el resto de 1. O sea, estas varianza se determinan así:

- var(1° componente principal) = var(z)+0.54603 = 1 + 0.54603 = 1.54603 = λ_1
- var(2° componente principal) = var(z)-0.54603 = 1 – 0.54603 = 0.45397 = λ_2

Por algebra lineal sabemos que cada autovalor tiene asociado un autovector. En el caso de 2 variables supongamos que estos autovectores tienen la forma de

$$\mathbf{u_1} = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} \text{ y } \mathbf{u_2} = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix}$$

sujeta a:

$$u_{11}^2 + u_{12}^2 = 1$$

$$u_{21}^2 + u_{22}^2 = 1$$

y que se cumple

$$\mathbf{u_1^t}\ \mathbf{u_2} = 0$$

O sea, son unitarios y ortogonales, estos vectores se llaman ortonormales o se denominan bases ortonormales.

Cuando los datos están tipificados los vectores que se obtienen, independientemente de los valores que tengan los autovalores, son los siguientes:

$$\mathbf{u_1} = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix} \text{ y } \mathbf{u_2} = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

como puede comprobarse, estos vectores cumplen con las restricciones dadas:

$$(0.7071)^2 + (\pm 0.7071)^2 = 1$$

y

$$\mathbf{u_1^t\ u_2} = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}^{\boldsymbol{t}} \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

$$= [\ 0.7071, 0.7071] \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

$$= 0$$

Los coeficientes o los elementos de los vectores $\mathbf{u_1}$ y $\mathbf{u_2}$ son los coeficientes que hay aplicar a las variables originales para obtener las respectivas componentes principales. Así:

$$Z_1 = u_{11}X_1 + u_{12}X_2$$
$$Z_2 = u_{21}X_1 + u_{22}X_2$$

En nuestro caso, como en todos los casos de dos variables tipificadas, las combinaciones lineales para la obtención de componentes son los siguientes:

$$Z_1 = 0.7071X_1 + 0.7071X_2$$
$$Z_2 = -\ 0.7071X_1 + 0.7071X_2$$

O sea,

$$Z_1 = 0.7071(1.257) + 0.7071(0.556) = 1.282$$
$$Z_2 = -\ 0.7071(1.257) + 0.7071(0.556) = -\ 0.496$$

luego tenemos la siguiente tabla.

Tabla 11: Componentes principales 1 y 2

N°	EMPRESAS	CP1	CP2
1	El Corte Inglés	1.282	-0.496
2	Iberdrola	2.253	0.473
3	Repsol Comercial	0.270	-0.717

4	Seat	-0.270	-0.969
5	Tabacalera	0.262	0.023
6	FASA Renault	-0.393	0.364
7	Repsol Petróleo	-0.556	0.720
8	Pryca	-0.747	0.963
9	Iberia	-2.100	-0.362

Los coeficientes de combinaciones lineales de Z_1 y Z_2 son precisamente los senos y los cosenos del ángulo de rotación entre los ejes de las componentes principales y los ejes correspondientes a las variables originales. Cuando se trata de variables tipificadas el ángulo de rotación es siempre de 45°. Así:

Primer eje principal: $\cos 45° = 0.7071$, $\text{sen}45° = 0.7071$

Segundo eje principal: $\cos 135° = -0.7071$, $\text{sen}135° = 0.7071$.

2.4 OBTENCIÓN DE LAS COMPONENTES PRINCIPALES EN EL CASO GENERAL DE n-VARIABLES.

Como sabemos que el análisis de componentes principales se concibe como una técnica de reducción de la dimensión pues permite pasar de una gran cantidad de variables intercorrelacionadas a unas pocas componentes principales. El método consiste en buscar combinaciones lineales de las variables originales que representen lo mejor posible a la variabilidad presente en los datos. De este modo, con unas pocas combinaciones lineales, que serán las componentes principales, sería suficiente para entender la información contenida en los datos. Al mismo tiempo, la forma en que se construyen las componentes, y su relación con unas u otras variables originales, sirven para entender la estructura de correlación inherente a los datos. Por último, las componentes principales, que forman un vector aleatorio de dimensión menor, pueden ser empleadas en análisis estadísticos posteriores.

2.4.1 REDUCCIÓN DE VARIABLES DE UN VECTOR ALEATORIO

La reducción de variables de un vector aleatorio se obtienen a partir de la matriz de covarianza de un vector aleatorio o de un conjunto de datos con la condición de que exista la matriz de covarianza.

Definición 2.1 Sea $X = (X_1, X_2, \ldots, X_n)^t$ un vector aleatorio n-dimensional con vector de medias $\mu = E[X]$ y matriz de covarianza $\Sigma = Cov(X, X)$. Se define la primera componente principal de X como una variable aleatoria Z_1 tal que

$$Z_1 = v_1^t X = v_{11}X_1 + v_{21}X_2 + \ldots + v_{n1}X_n \text{ con } v_1 = (v_{11}, v_{21}, \ldots, v_{n1})^t \in R^n$$

$$Var(Z_1) = \max\{var(v^tX);\ v \in R^n,\ v^tv = 1\}.$$

La primera componente principal es unas combinaciones lineales normalizadas de las variables de X y, de entre todas las combinaciones lineales normalizadas, es la que tiene mayor varianza.

Teorema 2.2 La primera componente principal de X adopta la forma

$$Z_1 = v_1^t X$$

Siendo λ_1 el mayor autovalor de Σ y v_1 un autovector asociado a λ_1 de norma 1. Además

$$Var(Z_1) = \lambda_1$$

Demostración.

Consideremos

$$Z = v^tX.$$

Entonces podemos calcular su varianza

$$Var(Z) = var(v^tX) = Cov(v^tX, v^tX) = v^tCov(X,X)v = v^t\Sigma v$$

Nuestro problema consiste en

$$\text{Max } (v^t\Sigma v)$$

$$\text{Sujeta a } v^tv = 1$$

Resolvemos este problema por el método de los multiplicadores de Lagrange

$$L = v^t\Sigma v - \lambda(v^tv - 1)$$

Calculamos la matriz Jacobiana de L como una función de v e igualamos a cero

$$\frac{\partial L}{\partial v} = 2v^t\Sigma - 2\lambda v^t = 0$$

Equivalentemente

$$\Sigma v = \lambda v$$

Luego el vector v_1 que maximice la función sujeta a la restricción ha de ser un autovector de Σ, y su autovalor asociado es λ_1. Multiplicando los dos términos de la ecuación por la izquierda por v_1^t resulta

$$v_1^t\Sigma v_1 = v_1^t\lambda_1 v_1 = \lambda_1 v_1^t v_1 = \lambda_1$$

Luego el autovalor λ_1 es la función objetivo en el máximo y por tanto

$$Var(Z_1) = \lambda_1$$

Definición 2.2 Se define la segunda componente principal de X como una variable aleatoria Z_2 tal que

$$Z_2 = v_2^t X = v_{12}X_1 + v_{22}X_2 + \ldots + v_{n2}X_n \text{ con } v_2 = (v_{12}, v_{22}, \ldots, v_{n2})^t \in R^n$$

$$Var(Z_2) = \max\{Var(v^tX);\ v \in R^n,\ v^tv = 1,\ v^tv=1,\ v^tv_1=0\ \}.$$

La segunda componente es otra combinación lineal de las variables de X y, de entre todas las combinaciones lineales formadas por vectores unitarios ortogonales a v_1, es la que tiene mayor varianza

Observación La siguiente igualdad

$$Cov(v^tX, Z_1) = Cov(v^tX, v_1^tX) = v^tCov(X, X)v_1 = v^t\Sigma v_1.$$

Proposición 2.3 Sea $w \in R^p$ un autovector de Σ asociado a un autovalor λ no nulo ($\lambda \neq 0$), sea $v \in R^n$. Entonces

$$v^t\Sigma w = 0 \Leftrightarrow v^tw = 0.$$

Demostración.

($\Leftarrow$) Se tiene

$$0 = v^t\Sigma w = v^t\lambda w = \lambda v^tw \Rightarrow v^tw = 0, \text{ pues } \lambda \neq 0.$$

($\Rightarrow$) Se tiene

$$0 = v^tw = v^t\frac{\lambda w}{\lambda} = v^t\frac{1}{\lambda}\Sigma w \Rightarrow v^t\Sigma w = 0.$$

Por tanto, en aplicación de este resultado

$$Cov(v^tX, Z_1) = 0 \Leftrightarrow v^tv_1 = 0.$$

El hecho de que

$$v^tv_1 = 0$$

significa que los vectores v_1 y v_2 son ortogonales, si y solo si,

$$Cov(v^tX, Z_1) = 0,$$

o sea, las componentes Z_1 y Z_2 son incorrelacionadas. Es así que, podríamos definir que la segunda componente principal es la combinación lineal de las variables de X que tiene mayor varianza de entre las combinaciones lineales normalizadas e incorrelacionadas con la primera componente principal.

El hecho de que $\lambda_1 \neq 0$, y esto se define como

$$\lambda_1 = Var(Z_1) \neq 0$$

Eso implicaría que si

$$\lambda_1 = Var(Z_1) = 0$$

Todos los demás autovalores serán también cero y Z_1 y las demás componentes serán variables autogenerados.

Teorema 2.4 La segunda componente principal de X adopta la forma

$$Z_2 = v_2^t X$$

Siendo λ_2 el segundo mayor autovalor de Σ y v_2 un autovector asociado a λ_2 de norma 1 ($v_2^t v_2 = 1$) y ortogonal a v_1 ($v_1^t v_2 = 0$). Además

$$Var(Z_2) = \lambda_2$$

Demostración.

La demostración es análoga al caso anterior. Consideremos

$$Z = v^t X$$

cuya varianza es

$$Var(Z) = v^t \Sigma v$$

Ahora el problema consiste en

$$\text{Max } v^t \Sigma v$$

$$\text{Sujeta a } v^t v = 1$$

$$v_1^t v = 0$$

Resolviendo este problema por el método de los multiplicadores de Langrage

$$L = v^t \Sigma v - \lambda_2 (v^t v - 1) - \mu_2 v_1^t v$$

Derivamos respecto de v e igualamos a cero (momentos en el cual podemos sustituir el valor v_2 para mejor compresión):

$$\left(\frac{\partial L}{\partial v}\Big|_{v=v_2}\right)^t = 2\Sigma v_2 - 2\lambda_2 v_2 - \mu_2 v_1 = 0 \quad (*)$$

Multiplicando por la izquierda por v_1^t resultan

$$2v_1^t\Sigma v_2 - 2\lambda_2 v_1^t v_2 - \mu_2 v_1^t v_1 = 0$$

Las restricciones imponen que

$$v_1^t v_2 = 0$$

Se deduce también que

$$v_1^t\Sigma v_2 = 0.$$

Y v_1 verificaba

$$v_1^t v_1 = 1.$$

Entonces la ecuación anterior se reduce a $\mu_2 = 0$. Y al sustituir ese valor en la ecuación (*)

$$\Sigma v_2 = \lambda_2 v_2$$

Razonando igual que antes, el vector v_2 que maximice la función sujeto a las restricciones ha de ser un autovector de Σ, y su autovalor asociado es λ_2. La restricción de ortonormalidad respecto de v_1 nos obliga a tomar el segundo mayor autovalor de Σ. Como antes,

$$\mathrm{Var}(Z_2) = v_2^t\Sigma v_2 = \lambda_2$$

Podemos continuar este proceso extrayendo las componentes principales de X mediante los autovalores de la matriz de covarianza Σ y la base ortonormal de autovectores asociado. De este modo, obtenemos n componentes principales.

Definición 2.3 Se definen las n componentes principales de X como las variables aleatorias $(Z_1, Z_2, \ldots, Z_n)$ tales que

$$Z_1 = v_1^t X,\ Z_2 = v_2^t X,\ \ldots,\ Z_n = v_n^t X\ ;\ v_1, v_2, \ldots, v_n \in R^n$$

$$\mathrm{Var}(Z_1) = \max\{\mathrm{Var}(v^tX): v\in R^n,\ v^tv = 1\}$$

$$\mathrm{Var}(Z_2) = \max\{\mathrm{Var}(v^tX): v\in R^n,\ v^tv = 1,\ v_1^tv = 0\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\mathrm{Var}(Z_n) = \max\{\mathrm{Var}(v^tX): v\in R^n,\ v^tv = 1,\ v_1^tv = 0,\ v_1^tv = 0,\ v_2^tv = 0,\ldots,\ v_{p-1}^tv = 0\}$$

Teorema 2.5 Las n componentes principales de X adoptan la forma

$$Z_j = v_j^t X,\ j \in \{1, 2, \ldots, n\}$$

Siendo $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$ los n autovalores ordenados de Σ y $v_1, v_2, \ldots, v_n$ sus autovectores asociados normalizados, esto es, $\{v_1, v_2, \ldots, v_n\}$ es una base ortonormal de autovectores. Además las componentes son incorreladas

$$Cov(Z_j, Z_k) = 0 \text{ si } j \neq k$$

y

$$Var(Z_j) = \lambda_j \, , \, j \in \{1, 2, \ldots, n\}$$

Demostración.

Supongamos

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0.$$

Probemos que las variables

$$Z_j = v_j^t X, \, j \in \{1, 2, \ldots, n\}$$

son incorrelacionadas:

$$Cov(Z_i, Z_j) = v_i^t \Sigma v_j = v_i^t \lambda_j v_j = \lambda_j v_i^t v_j$$

$$Cov(Z_j, Z_i) = v_j^t \Sigma v_i = v_j^t \lambda_i v_i = \lambda_i v_j^t v_i$$

Entonces

$$(\lambda_j - \lambda_i) \, v_i^t v_j = 0,$$

$$v_i^t v_j = 0 \Rightarrow Cov(Z_i, Z_j) = v_i^t \Sigma v_j = v_i^t \lambda_j v_j = \lambda_j v_i^t v_j = 0, \, i \neq j.$$

Además:

$$Var(Z_j) = \lambda_j v_j^t v_j = \lambda_j \, , \, j \in \{1, 2, \ldots, n\}$$

Sea ahora

$$Z = \sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \alpha_i Z_i$$

una variable compuesta tal que

$$\sum_{i=1}^{n} \alpha_i^2 = 1$$

$$\begin{aligned} Var(Z) &= Var\left(\sum_{i=1}^{n} a_i X_i\right) \\ &= Var\left(\sum_{i=1}^{n} \alpha_i Z_i\right) \\ &= \sum_{i=1}^{n} \alpha_i^2 Var(Z_i) \\ &= \sum_{i=1}^{n} \alpha_i^2 \lambda_i \\ &\leq \left(\sum_{i=1}^{n} \alpha_i^2\right) \lambda_1 \\ &= Var(Z_1) \end{aligned}$$

Que prueba que Z_1 tiene varianza máxima.

Consideremos ahora las variables Z incorrelacionadas con Z_1. Las podemos expresar como:

$$Z = \sum_{i=1}^{n} b_i X_i = \sum_{i=1}^{n} \beta_i Z_i$$

condicionada a

$$\sum_{i=1}^{n} \beta_i^2 = 1$$

Entonces

$$\begin{aligned} Var(Z) &= Var(\sum_{i=1}^{n} \beta_i Z_i) \\ &= \sum_{i=1}^{n} \beta_i^2 \, Var(Z_i) \\ &= \sum_{i=1}^{n} \beta_i^2 \lambda_i \\ &\leq (\sum_{i=1}^{n} \beta_i^2) \lambda_2 \\ &= Var(Z_2), \end{aligned}$$

Y por tanto Z_2 está incorrelacionadas con Z_1 y tiene varianza máxima. Si $n \geq 3$, la demostración de que

$$Z_3, \; Z_4, \ldots, Z_n$$

Son también componentes principales es análoga.

En base a los resultados obtenidos, podemos resumir que para un vector aleatorio

$$X^t = (X_1, X_2, \ldots, X_n)$$

Tiene una matriz de varianza-covarianzas Σ. Sin pérdida de generalidad asumir que la media de los X_i es cero, para todo $i = 1, 2, \ldots, n$; esto siempre lícito, pues de otra manera sólo basta con centrar (restando la media) el vector X. Como se dijo anteriormente que, para encontrar la primera componente principal, se examina el vector de coeficientes

$$v_1^t = (\lambda_{11}, \; \lambda_{12}, \ldots, \; \lambda_{1n})$$

tal que, la varianza

$$v_1^t X$$

sea un máximo con la restricción $v_1^t v_1 = 1$. De esta manera, se determina la combinación lineal

$$Z_k = \lambda_{k1} X_1 + \lambda_{k1} X_2 + \ldots + \lambda_{kn} X_n = \sum_{j=1}^{n} \lambda_{kj} X_j \; ; \; k = 1, 2, \ldots, r \leq n$$

tal que

$$Var(Z_k) = Var\left(\sum_{j=1}^{n} \lambda_{kj} X_j \right)$$

sea máxima, donde

$$\sum_{j=1}^{n} \lambda_{kj}^2 = 1$$

Para k = 1, 2, …, r ≤ n. Y que $v_i^t v_j = 0$, para todo i ≠ j.

Ejemplo 2.2 Supongamos que deseamos conocer cuales son los factores relacionados con el riesgo de enfermedad coronaria. Del conocimiento previo sabemos que el riesgo es la presión arterial, la edad, la obesidad, el tiempo que se ha sido hipertenso, el pulso y el stress. Para la investigación seleccionado al azar 20 pacientes hipertenso en los que medimos las siguientes variables:

X_1: Presión arterial media (mm Hg).

X_2: Edad (años)

X_3: Peso (Kg)

X_4: Superficie corporal (m^2)

X_5: Duración de la Hipertenso (años)

X_6: Pulso (pulsaciones/minutos)

X_7: Medida del stress.

Tratamos de estudiar la situación del grupo de pacientes en relación a los factores de riesgo y las posibles interrelaciones entre las distintas variables. Inicialmente queremos describir el conjunto de pacientes utilizando simultáneamente todas las variables.

Los datos obtenidos se muestran en la tabla siguiente.

Tabla: 12: Presentación de datos de enfermedad coronaria

Nº	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	46	95.7	2.09	7.0	75	99

En este problema estamos trabajando con 7 variables, o sea, la dimensión es 7, pero ¿ es posible describir el conjunto de datos utilizando con un número menor de dimensiones, aprovechando las interrelaciones entre las variables?. Según la teoría desarrollada en esta tesis, la respuesta es afirmativa en el sentido de que; si es posible describir el problema con un número menor de dimensión. Las interrelaciones entre variables quedan reflejadas en la matriz de covarianza Σ, el procedimiento en general será obtenida mediante el programa estadístico SPSS versión 22.

En primer lugar vamos a estudiar la existencia de las siete variables de dos en dos. Mediante el programa SPSS version 22 se tiene:

Tabla 13: Matriz de correlaciones de enfermedad coronaria

		X1	X2	X3	X4	X5	X6	X7
Correlación	X1	1,000	,659	,950	,866	,293	,721	,164
	X2	,659	1,000	,407	,378	,344	,619	,368
	X3	,950	,407	1,000	,875	,201	,659	,034
	X4	,866	,378	,875	1,000	,131	,465	,018
	X5	,293	,344	,201	,131	1,000	,402	,312
	X6	,721	,619	,659	,465	,402	1,000	,506
	X7	,164	,368	,034	,018	,312	,506	1,000

En la Tabla 13 se observa que la correlación de Pearson r es como sigue:

$r_1 = Corr(X_1, X_2) = 0.659 = r_{12}$

$r_2 = Corr(X_1, X_3) = 0.950 = r_{13}$

$r_3 = Corr(X_1, X_4) = 0.866 = r_{14}$

$r_4 = Corr(X_1, X_5) = 0.293 = r_{15}$

$r_5 = Corr(X_1, X_6) = = 0.721 = r_{16}$

$r_6 = Corr(X_1, X_7) = 0.164 = r_{17}$

$r_7 = Corr(X_2, X_3) = 0.407 = r_{23}$

$r_8 = Corr(X_2, X_4) = 0.0.378 = r_{24}$

$r_9 = Corr(X_2, X_5) = 0.344 = r_{25}$

$r_{10} = Corr(X_2, X_6) = 0.619 = r_{26}$

$r_{11} = Corr(X_2, X_7) = 0.368 = r_{27}$

$r_{12} = Corr(X_3, X_4) = 0.875 = r_{34}$

$r_{13} = Corr(X_3, X_5) = 0.201 = r_{35}$

$r_{14} = Corr(X_3, X_6) = 0.659 = r_{36}$

$r_{15} = Corr(X_3, X_7) = 0.034 = r_{37}$

$r_{16} = Corr(X_4, X_5) = 0.131 = r_{45}$

$r_{17} = Corr(X_4, X_6) = 0.465 = r_{46}$

$r_{18} = Corr(X_4, X_7) = 0.018 = r_{47}$

$r_{19} = Corr(X_5, X_6) = 0.402 = r_{56}$

$r_{20} = Corr(X_5, X_7) = 0.312 = r_{57}$

$r_{21} = Corr(X_6, X_7) = 0.506 = r_{67}$

Hay una relación matemática que cumple cuando se relacionan dos variable en dos variables, o sea, tenemos

$$\binom{7}{2} = \frac{7!}{2!5!} = 21$$

correlacionamientos.

En el análisis anterior tenemos las siguientes correlaciones altas:

$$\mathbf{r_2 = Corr(X_1, X_3) = 0.950 = r_{13}}$$

$$\mathbf{r_3 = Corr(X_1, X_4) = 0.866 = r_{14}}$$

$$\mathbf{r_{12} = Corr(X_3, X_4) = 0.875 = r_{34}}$$

significa que la variable X_1 con X_3 tienen una correlación de 95%, o sea, esas variables están fuertemente correlacionadas, lo cual significa que la presión arterial esta fuertemente correlacionada con el peso del individuo.

Las variables X_1 y X_4 su correlación equivale a 86.6% también están fuertemente correlacionadas. Lo cual indica la presión arterial estan fuertemente correlacionadas con la superficie corporal del individuo.

De la misma forma observamos que las variables X_3 con X_4 se tiene una una correlación de 87.5%, lo cuál indica que el sobrepeso de una persona esta fuertemente correlacionada con la superficie corporal.

Con respecto a las correlaciones bajas podemos decir.

$$\mathbf{r_{15} = Corr(X_3, X_7) = 0.034 = r_{37}}$$

$$\mathbf{r_{18} = Corr(X_4, X_7) = 0.018 = r_{47}}$$

Tenemos que las variables X_3 con X_4 tienen una correlación de 3.4%, eso significa que el peso de una persona no tiene mucha correlación con el stresamiento del individuo.

De la misma forma podemos deducir que las variables X_4 y X_7 con 1.8% de correlacionamiento, o sea, la superficie corporal de una persona no hay tanta correlación con el stresamiento de dicha persona.

Tabla 13: Varianza total explicada

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	3,908	55,833	55,833	3,908	55,833	55,833
2	1,470	21,003	76,836	1,470	21,003	76,836
3	,709	10,126	86,961			
4	,522	7,453	94,414			
5	,308	4,399	98,814			
6	,081	1,155	99,968			
7	,002	,032	100,000			

A partir de esta tabla podemos decir que la varianza total VT es

$$VT = 3.908 + 1.470 + 0.709 + 0.522 + 0.308 + 0.081 + 0.002 = 7$$

entonces

$$7 \rightarrow 100\%$$

$$3.908 \rightarrow x\%$$

de modo que:

$$x1\% = \frac{3.908 \times 100\%}{7} = 55.83\%$$

Del mismo modo se tiene:

$$7 \rightarrow 100\%$$

$$1.470 \rightarrow x2\%$$

De modo que:

$$x2\% = \frac{1.470 \times 100\%}{7} = 21\%$$

sumando los dos resultados se tiene:

$$x1\% + x2\% = 21\% + 55.83\% = 76.83\% \cong 77\%$$

En este caso, las dos primeras componentes recogen aproximadamente el 77% de la variabilidad, más aun recogen las fuentes de variabilidad más importantes de los datos. El resto de los datos ofrecen solamente 33% de variabilidad de información aproximadamente. Una vez reducidos la cantidad de variables originales participantes, obtenemos ciertas variables nuevas llamada componentes. Como las componentes son variables compuestas calculadas a partir de las originales, solamente queda por determinar cual es la información que han recogido las componentes, es decir, que variables explican la similitud de los inidividuos en el subespacio de representación final. La interpretación se hace a partir de las

correlaciones entre las variables observadas y las componentes. Dichas correlaciones se muestran en la tabla siguiente. (las componentes también se llaman factores en la tabla).

Tabla 14: Matriz de componentes

	Componente	
	1	2
Presión: X1	,965	-,230
Edad : X2	,723	,304
Peso : X3	,884	-,403
Supcor : X4	,804	-,473
Durac : X5	,434	,525
Pulso : X6	,844	,284
Stress : X7	,355	,764

Observamos como la primera componente está altamente correlacionada con todas las variables salvo Duración y Stress, es decir, la primera componente muestra, fundamentalmente aspectos relacionados con el aumento de la presión arterial y de las variables determinantes del riesgo de enfermedad coronaria.

La segunda componente está más correlacionada con el stress y algo menos con la duración, por lo que mostrará las diferencias en el índice de stress.

CAPÍTULO III

APLICACIÓN A LA INDUSTRIA MANUFACTURERA.

3.1 INTRODUCCIÓN.

Para poder aplicar al teoría y la metodología desarrollada en el capítulo II, utilizaré la información publicada en la página web del Instituto Nacional de Estadística e Informática del Perú (INEI), considerando de antemano que dichas encuestas son elaboradas por los expertos del INEI, como también la captación de datos.

3.2 MUESTRA ADECUADA DE LA ENCUESTA DE INNOVACIÓN EN LA INDUSTRIA MANUFACTURERA

El Instituto Nacional de Estadística e Informática (INEI) del estado peruano, realiza anualmente encuestas en los sectores productivos como Industria Manufacturera, Población Económicamente Activa, Producción Agropecuaria, Encuesta Nacional de Hogares, etc. Para mi tesis utilizaré como aplicación LA ENCUESTA DE INNOVACIÓN EN LA INDUSTRIA MANUFACTURERA del año 2009-2011. Dichas encuesta tienen su ficha técnica y el instrumento correspondiente, que en este caso es una encuesta, los cuales aparecerán en el anexo respectivo.

Las variables con las que se va trabajar son:

1. Y1: Fecha de resultado final de la encuesta_DIA (FECHA FINAL-P2_RESFIN_DIA)
2. Y2: Fecha de resultado final de la encuesta_AÑO (AÑO FINAL-P2_RESFIN_ANIO)
3. Y3: Distribuya porcentualmente los fondos de financiamiento utilizados por la empresa durante el periodo 2009 - 2011 para la realización de actividades de innovación, según el origen de los mismos: Fuentes propias - Recursos propios. (FUENTES PROPIOS-RECURSO PROPIOS-P4_1_1)

4. Y4: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Postgrado?. (PERSONAL CON POSGRADO-P5_1_1_2011)
5. Y5: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Superior Universitaria Completa? (PERSONAL UNIVERSITARIO COMPLETO-P5_1_2_2011)
6. Y6: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Superior Universitaria Incompleta? (PERSONAL UNIVERSITARIO INCOMPLETO-P5_1_3_2011)
7. Y7: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Superior No Universitaria Completa? (PERSONAL TÉCNICO NO UNVERSITARIO-P5_1_4_2011)
8. Y8: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Superior No Universitaria Incompleta? (PERSONAL TÉCNICO NO UNIVERSITARIA INCOMPLETO-P5_1_5_2011).
9. Y9: En el año 2009 y 2011, según el último nivel de estudios alcanzado ¿cuál fue el número promedio del personal ocupado con Secundaria Completa? (PERSONAL SECUNDARIA COMPLETA-P5_1_6_2011)
10. Y10: Promedio total del personal ocupado en el año 2009 (PROMEDIO PERSONAL 2009-P5_1_13_2009)
11. Y11: Promedio total del personal ocupado en el año 2011 (PROMEDIO PERSONAL 2011-P5_1_13_2011)
12. Y12: ¿Cuál fue el número de trabajadores en el año 2011 con estudios de postgrados y superior universitaria concluidos con formación en: Ciencias sociales? (PROFESIONAL EN CIENCIAS SOCIALES-P5_2_5)
13. Y13: Total (TOTAL PERSONAL-P5_2_8)
14. Y14: ¿Cuál es la cantidad promedio de trabajadores que en el año 2011, se desempeñaron en el área Funcional de: Informática y Sistemas? (PERSONAL DE INFORMÁTICA Y SISTEMAS-P5_3_1).
15. Y15: Anote el número del ítem de innovación más importante que haya introducido la empresa en el periodo 2009 - 2011, ya sea en producto, proceso, organización o comercialización (PRODUCCIÓN DE LA EMPRESA-P6_3).

16. Y16: Distribuya porcentualmente en forma vertical el valor de las ventas al mercado interno del año 2011 según el grado de novedad de la innovación de producto (bien o servicio) en el periodo 2009 - 2011: nuevos o significativamente mejorados tanto para la empresa (CANTIDAD PORCENTUAL DE LA INNOVACIÓN-P7_1_1_INT)
17. Y17: Distribuya porcentualmente en forma vertical el valor de las ventas al mercado externo del año 2011 según el grado de novedad de la innovación de producto (bien o servicio) en el periodo 2009 - 2011: Iguales o que no fueron alterados significativamente los porcentuales de producción. (DISTRIBUCIÓN PORCENTUAL DE LA INNOVACIÓN DE PRODUCTO-P7_1_1_EXT).
18. Y18: ¿Cuál fue el desempeño, en cuanto al porcentaje promedio de utilización de capacidad instalada el año 2011? (PORCENTAJE PROMEDIO DE CAPACIDAD-P12_3_4_2011)
19. Y19: Factor de expansión. (FACTOR DE EXPANCIÓN-facexp)

Según el análisis muestral, hemos considerado 293 como tamaño de la muestras. El programa estadístico SPSS-22 nos ofrece en primer lugar los Estadísticos Descriptivos esto es importante porque aparece el tamaño de la muestral con lo que estamos trabajando, aparte de que nos muestra la media y la desviación estándar de las variables participantes. Tal como podemos observar en la Tabla 15. Después de realizar los estadísticos descriptivos vamos a pasar ver las correlaciones que tiene entre variables participantes. Tal como se muestral en la Tabla 16.

Tabla 15: Estadísticos descriptivos de recursos manufactorados

	Media	Desviación estándar	N de análisis
Y3: FUENTES PROPIOS-RECURSOS PROPIOS	66,01	33,160	293
Y4: PERSONAL CON POSGRADO	5,96	18,026	293
Y5: PERSONAL UNIVERSITARIO COMPLETO	40,77	113,127	293
Y6: PERSONAL UNIVERSITARIO INCOMPLETO	10,59	30,650	293
Y7: PERSONAL TÉCNICO NO UNIVERSITARIO	50,66	149,917	293
Y8: PERSONAL TÉCNICO NO UNIVERSITARIA INCOMPLETA	15,97	56,330	293
Y9: PERSONAL CON SECUNDARIA COMPLETA	141,89	406,904	293
Y10: PROMEDIO DE CANTIDAD PERSONAL-2009	285,53	725,673	293
Y11: PROMEDIO DE CANTIDAD PERSONAL-2011	317,49	830,858	293
Y12: PROFESIONAL EN CIENCIAS SOCIALES	17,47	54,503	293
Y13: TOTAL PERSONAL	46,73	123,175	293
Y14: PERSONAL DE INFORMÁTICA Y SISTEMAS	3,37	11,380	293
Y15: PRODUCCIÓN DE LA EMPRESA	1,720	,4979	293
Y16: CANTIDAD PORCENTUAL DE LA INNOVACIÓN	24,11	34,469	293
Y17: DISTRIBUCIÓN PORCENTUAL DE LA INNOVACIÓN DE PRODUCTO	30,39	40,486	293
Y18: PORCENTAJE PROMEDIO DE CAPACIDAD	71,31	19,609	293

En esta Tabla 16 mostramos las correlaciones que existen entre variables participantes, dos variables altamente correlacionadas se puede tratar como una sola variable cuando queremos aplicar la reducción de dimensionalidad.

Tabla 16: Matriz de correlaciones de recursos manufactorados

	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15
00	,128	,105	,024	,120	,091	,015	,103	,098	,088	,115	-,010	,094
28	1,000	,502	,126	,583	,199	,435	,539	,613	,529	,608	,236	-,041
05	,502	1,000	,272	,592	,262	,380	,535	,541	,664	,992	,384	-,068
24	,126	,272	1,000	,332	,472	,304	,444	,392	,233	,268	,334	-,005
20	,583	,592	,332	1,000	,329	,379	,549	,584	,592	,629	,340	-,041
91	,199	,262	,472	,329	1,000	,246	,464	,376	,132	,270	,133	-,001
15	,435	,380	,304	,379	,246	1,000	,789	,852	,242	,412	,286	-,039
03	,539	,535	,444	,549	,464	,789	1,000	,961	,386	,570	,338	-,022
98	,613	,541	,392	,584	,376	,852	,961	1,000	,400	,586	,314	-,025
88	,529	,664	,233	,592	,132	,242	,386	,400	1,000	,687	,192	-,034
15	,608	,992	,268	,629	,270	,412	,570	,586	,687	1,000	,387	-,069
10	,236	,384	,334	,340	,133	,286	,338	,314	,192	,387	1,000	-,012
94	-,041	-,068	-,005	-,041	-,001	-,039	-,022	-,025	-,034	-,069	-,012	1,000
89	,117	,057	,042	,044	-,002	-,009	-,008	,000	,173	,070	,020	-,223
77	,166	,221	,181	,244	,181	,249	,268	,265	,154	,227	,167	-,121
05	,137	,111	,039	,104	,058	,039	,061	,060	,138	,122	,106	,112

En la Tabla 16 se tiene:

- Desde un punto de vista observamos de manera general que algunas variables se asocian cuyos puntajes de correlacionamiento sobre-pasan 0.5.
- Tenemos Corr(Y4, Y11) = 0.613, indica que tienen un 61.3% de correlación, o sea, según la instrumento podemos decir que en el año 2009 y 2011, según el último nivel de estudio alcanzado, ¿cuál fue el número promedio del personal ocupado con Posgrado? se correlaciona en un 61.3% con el Promedio total del personal ocupado en el año 2009.
- Tenemos Corr(Y4, Y13) = 0.608, indica que tienen un 60.8% de correlación, o sea, según la instrumento podemos decir que. En el año 2009 y 2011, según el último nivel de estudio alcanzado, ¿cuál fue el número promedio del personal ocupado con Posgrado? se correlaciona en un 60.8% con el total personal.
- Se observa también Corr(Y7, Y13) = 0.629, nos indica un 62.9% de correlacionamiento entre en número promedio del personal ocupado con Educación Superior no Universitario Completo y el total personal.
- La Corr(Y12, Y13) = 0.687, indica que hay un 68.7% de correlacionamiento, o sea, entre el número de trabajadores en el año 2011 con estudios de posgrado con formación en Ciencias Sociales y total personal.

Son eso los casos que podemos anotar viendo el nivel de correlacionamiento entre ellos.

En la Tabla 17 que es lo importante, porque mide la varianza máxima mediante la obtención de autovalores, así tenemos que:

1. La primera nueva variable retiene una variabilidad de 36.013%.
2. La segunda nueva variable retiene una variabilidad de 9.979%.
3. La tercera nueva variable retiene una variabilidad de 7.898%.
4. La cuarta nueva variable retiene una variabilidad de 7.345%
5. La quinta nueva variable retiene una variabilidad de 6.793%.
6. La sexta nueva variable retiene una variabilidad de 6.303%

Según el análisis estadístico de SPSS, sólo es importante analizar 6 de tales variables originales.

Tabla 17: Varianza total explicada de recursos manufactorados

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	5,762	36,013	36,013	5,762	36,013	36,013
2	1,597	9,979	45,993	1,597	9,979	45,993
3	1,264	7,898	53,891	1,264	7,898	53,891
4	1,175	7,345	61,235	1,175	7,345	61,235
5	1,087	6,793	68,028	1,087	6,793	68,028
6	1,008	6,303	74,331	1,008	6,303	74,331
7	,911	5,694	80,024			
8	,749	4,682	84,706			
9	,575	3,595	88,301			
10	,557	3,482	91,783			
11	,479	2,996	94,780			
12	,354	2,214	96,994			
13	,285	1,781	98,775			
14	,174	1,086	99,861			
15	,022	,139	100,000			
16	2,851E-16	1,782E-15	100,000			

Como hemos visto en la Tabla 17 que la reducción de variable ha sido 6 variables nuevas. Entonces es suficiente hacer un análisis del problema original con estudiar sólo 6 nuevas variables no correlacionadas. Dónde:

- La primera nueva variable casi no se relaciona nada con Y3, Y15 y Y16, directa o inversamente.
- La segunda nueva variable se relaciona de manera negativa con Y6, Y8 y Y9.
- La tercera nueva variable regularmente se relaciona con Y3, Y15 y Y18
- La cuarta nueva variable se relaciona inversa y directamente con Y15 y Y16 respectivamente.
- La quinta variable nueva no se relaciona casi nada con Y15 y Y18.
- La sexta nueva variable casi nada relaciona con Y4 y Y11.

Tabla 18: Matriz de coeficiente de puntuación de componente

	Componente					
	1	2	3	4	5	6
Y3	,026	,140	,503	,038	-,156	-,249
Y4	,124	,143	-,039	-,086	-,266	-,030
Y5	,140	,203	-,103	-,157	,183	,045
Y6	,085	-,218	,124	,302	,398	,247
Y7	,134	,093	-,012	-,064	,093	,026
Y8	,080	-,221	,202	,265	,254	,129
Y9	,121	-,251	-,055	,042	-,385	-,050
Y10	,149	-,214	,015	,058	-,231	,029
Y11	,151	-,188	-,017	,019	-,316	,001
Y12	,116	,300	-,078	-,107	,092	,078
Y13	,147	,208	-,100	-,157	,130	,037
Y14	,082	-,032	-,010	,023	,347	,124
Y15	-,011	-,072	,448	-,459	-,057	,475
Y16	,013	,309	-,008	,611	-,181	,144
Y17	,062	-,140	,057	-,097	,277	-,737
Y18	,028	,204	,496	,129	-,008	-,180

El proceso de reducción de variables en un análisis de multivariable, es un estudio de tipo descriptivo, ya que, no es un estudio predictivo ni exploratorio. Razones por las cuales, en este tipo de estudio lo que se persigue es que las nuevas variables halladas que expliquen la variabilidad de la información obtenida con las variables originales, sean no correlacionadas pero igual que las variables originales explique mayor cantidad de variabilidad que llevan las variables originales. Tal como se muestra en la tabla 19.

Tabla 19: Matriz de covarianzas de puntuación de componente

Componente	1	2	3	4	5	6
1	1,000	,000	,000	,000	,000	,000
2	,000	1,000	,000	,000	,000	,000
3	,000	,000	1,000	,000	,000	,000
4	,000	,000	,000	1,000	,000	,000
5	,000	,000	,000	,000	1,000	,000
6	,000	,000	,000	,000	,000	1,000

IV. CONCLUSIONES.

1. La técnica de reducción de variables, es una herramienta que debe usarse principalmente para una técnica exploratoria que ayuda a los investigadores de datos multivariantes que adquieran cierta percepción respecto a un conjunto de datos en el sentido de comprender mejor la estructura de correlación entre las variables.
2. La reducción de dimensionalidad en un análisis de datos multivariados, permite crear un espacio multidimensional menor que el espacio multidimensional original, permitiendo así determinar la dimensionalidad real de los datos con máxima variabilidad. Esto permitirá reemplazar a los datos originales por un número menor de variables subyacentes, sin que se pierda la información.
3. La aplicación del procedimiento de reducción de variables requiere básicamente en calcular los autovalores y autovectores asociado a la matriz de varianza-covarianza , de modo que la varianza de cada componente principal es igual al valor de la raíz del polinomio característico a que está asociado.

V. BIBLIOGRAFIA

1. **CUADRAS CHARLES M.** Nuevos Métodos de Análisis Multivariante. Edit. CMC Ediciones Barcelona 2010- España.

2. **DIAZ MONROY LUIS GUILLERMO / MORALES RIVERA MARIO ALFONSO.** Análisis Estadístico de Datos Multivarariados. Universidad Nacional de Colombia, Bogotá 2012.

3. **JOHNSON DALLAS E.** Métodos Multivariantes Aplicados al Análisis de datos. Edit. Thomson Editores, Mexico-2000

4. **JOHNSON RICHARD A.** Applied Multivariate Statistical Analysis. Edit, Prentice Hall, New Jersey- 1992-USA.

5. **LÉVY MANGIN JEAN-PIERRE Y VARELA MALLOU JESÚS**. Análisis Mulitvariable para las Ciencias Sociales Edit. Prentice Hall Madrid 2003- España.

6. **MARDIA K. V. KENT J. T. AND BIBBY J. M.** Multivariate Analysis. Edit. Academic Press 1979-Toronto USA.

7. **PEÑA DANIEL.** Análisis de Datos Multivariados. Mc-Graw Hill España 2002.

8. **PÉREZ CÉSAR.** Técnicas de Análisis Multivariante de Datos –Aplicaciones con SPSS. Edit. Pearson Prentice Hall. España-2004.

9. **URIEL EZEQUIEL, ALDAS JOAQUIN.** Análisis Multivariante Aplicado. Edit. THOMSON de México 2005.

10. **VISAUTA VINACUA BIENVENIDO.** Análisis Estadístico con SPSS Para Windows Edit Mc Graw Hill Madrid 2002-España.