ENFOQUES HISTÓRICOS DEL ANÁLISIS DE COMPONENTES PRINCIPALES Y EQUIVALENCIA ENTRE ELLOS.

Busto Guerrero, José Javier

jjbusto@cica.es

Romero García, José Enrique.

romerogje@yahoo.es

Departamento de Economía Aplicada I.

Facultad de Ciencias Económicas y Empresariales. Universidad de Sevilla.

Avda. Ramón y Cajal, Nº 1, 41005-Sevilla.

Área Temática: Métodos Cuantitativos

Resumen de la comunicación:

El Análisis de Componentes Principales (ACP) es una técnica antigua, los pioneros en el desarrollo de la teoría del ACP fueron Pearson (1901) y Hotelling (1933), si bien los términos de CP y ACP fueron acuñados por Hotelling.

Los enfoques de Pearson y Hotelling sobre la problemática planteada son diferentes. El enfoque de Pearson consistió en consideraciones sobre los puntos; mientras que el de Hotelling se baso en consideraciones sobre las variables. En este trabajo veremos ambos enfoques y la equivalencia entre ellos.

Palabras claves: Componentes Principales

1 Introducción.-

Los pioneros en el desarrollo de la teoría del ACP fueron Pearson (1901) y Hotelling (1933), si bien los términos de CP y ACP fueron acuñados por Hotelling.

Los enfoques de Pearson y Hotelling sobre la problemática planteada son diferentes.

El enfoque de Pearson consistió en trabajar sobre puntos, buscaba el hiperplano de mejor ajuste al sistema de puntos en estudio ; mientras que el de Hotelling se baso en el trabajo sobre las variables, buscaba las variables fundamentales que determinaban los valores de las variables originales. Analizaremos ambos enfoques, haremos demostraciones

alternativas a las realizadas por Pearson y Hotelling, y veremos que los dos enfoques coinciden.

Consideremos la matriz de datos $\mathbf{X}\in \hat{\mathbf{A}}^{nxp}$ centrada, $\mathbf{X}=(x_{i,j})$, $i=1,..,n$; $j=1,..,p$

$$\mathbf{X}=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Los datos que aparecen en la matriz **X** significan que tenemos n observaciones, $\mathbf{f_1}^t,..,\mathbf{f_n}^t$, de un *p*-vector $\mathbf{x=(x_1,...,x_p)}^t$. Es decir, las filas son medidas de los individuos u observaciones sobre *p* variables, con lo que $\mathbf{X=(f_1,..,f_n)}^t$, donde $\mathbf{f_i}^t=(x_{i1},..,x_{ip})\in \hat{\mathbf{A}}^p$, $i=1,..,n$; y las columnas representan los valores de las variables sobre los individuos, con lo que $\mathbf{X=(x_1,..,x_p)}$, donde $\mathbf{x_j}=(x_{1j},..,x_{nj})^t\in \hat{\mathbf{A}}^n$.

La matriz de varianzas-covarianzas, **S**, puede escribirse como

$$\mathbf{S}=\frac{1}{n}\mathbf{X}^t\mathbf{X},$$

2 ACP desde el enfoque de Hotelling.-

2.1 Obtención de las CP según Hotelling.-

Si consideramos *p* variables, $\mathbf{x_1,..,x_p}$, medidas sobre *n* individuos; por ejemplo, las paridades entre varias monedas; usualmente, las variables $\mathbf{x_i}$ estarán correladas. Es natural, dice Hotelling [1], que pueda haber un menor grupo fundamental de variables incorreladas (él las llama independientes) que determinen los valores de las *p*-variables originales. Si $\mathbf{z_1,..,z_p}$ son tales variables, tendremos un conjunto de relaciones de la forma

$$\mathbf{x_i}=f_i(\mathbf{z_1,..,z_p}),\ i=1,..,p. \qquad (\text{nosotros pondremos } \mathbf{z_i}=f(\mathbf{x_1,..,x_p}))$$

Hotelling usa el término *componentes* para referirse a las variables fundamentales encontradas. Restringiéndonos al caso en que las funciones f son lineales, tendremos que la expresión anterior queda en la forma:

$$\mathbf{x_i}=\Sigma_j\, a_{ij}\,\mathbf{z_j} \qquad (\text{nosotros pondremos } \mathbf{z_j}=\Sigma_j\, a_{ji}\,\mathbf{x_j})" \qquad (2.1.1)$$

Hotelling continúa diciendo, en otra parte de su trabajo, : " Un caso análogo es el uso de las ecuaciones de regresión que envuelven más y más variables $\mathbf{x_1,x_2,...}$ para explicar o predecir la variable **y**, ellas comienzan eligiéndose según el orden de sus contribuciones a la varianza de **y**. Esa analogía, sugiere que de las infinitas formas posibles que tenemos para obtener *componentes* a partir de las variables originales, comencemos con una componente $\mathbf{z_1}$, cuya contribución a las varianzas de las variables originales sea tan grande como sea posible; la próxima componente, $\mathbf{z_2}$, la tomamos incorrelada con $\mathbf{z_1}$, y con una contribución a la varianza residual tan grande como sea posible; y así seguimos procediendo para determinar las componentes, no excediéndonos en número de *p*, y puede que desechemos aquellas cuyas contribuciones a la varianza total sean pequeñas. Este procedimiento lo llamaremos *método de componentes principales*".

Hotelling es pues, el que introduce el término "componentes" para esas variables fundamentales y escoge dichas "componentes" de manera que maximicen sus sucesivas contribuciones al total de las varianzas de las variables originales, y llama a las componentes así obtenidas "Componentes Principales", y al procedimiento para hallarlas "Método de Componentes Principales " o "Análisis de Componentes Principales".

El procedimiento usado por Hotelling y el que desarrollaremos en este apartado son esencialmente los mismos, solo se diferencian en algunos aspectos: él consideraba las variables originales como funciones lineales de las Componentes, nosotros consideraremos las Componentes como funciones lineales de las variables originales lo que nos permitirá expresar la varianza de la Componente Principal a partir de la matriz de varianzas-covarianzas de las variables originales; él trabajaba en un contexto poblacional y consideraba solo el caso en que las componentes principales seguían una distribución normal y las consideraba de varianza unidad, nosotros trabajaremos en uno puramente descriptivo; él partía de variables tipificadas y llegaba a un problema de autovalores y autovectores de la matriz de correlación, nosotros partiremos de variables centradas con lo que llegaremos a un problema de autovectores y autovalores de la matriz de covarianzas; él no usaba notación matricial y nosotros sí.

Sea $\mathbf{x, x=(x_1,..,x_p),}$ el vector de *p* variables unidimensionales objeto de investigación; sea **S** la matriz de covarianzas de dicho vector.

Las componentes principales, tal como fueron determinadas inicialmente por Hotelling, en su trabajo original, se hallan como indicamos a continuación:

Hemos considerado, por (2.1.1), que $\mathbf{x_i}=\Sigma_j\, a_{ij}\, \mathbf{z_j}$

La correlación entre las variables originales *i* y *k*, r_{ik}, puede ponerse como la covarianza de las variables tipificadas; de donde:

$$r_{ik}=Cov(\mathbf{x_i}, \mathbf{x_k})= E(\mathbf{x_i}\,\mathbf{x_k})- E(\mathbf{x_i})\,E(\mathbf{x_k})= E(\mathbf{x_i}\,\mathbf{x_k}) = \text{Por (2)} = \sum_j a_{ij}\, a_{kj}\, E(\mathbf{z_j}^2) \qquad (2.1.2)$$

Pero hemos considerando que las CPs tenian media cero y varianza unidad, luego

$$1=Var(z_j)= E(\mathbf{z_j}^2) \qquad (2.1.3)$$

Por (2.1.2) y (2.1.3), se tiene que

$$r_{ik}= \Sigma_j\, a_{ij}\, a_{kj} \qquad (2.1.4)$$

Por (2.1.1), $\mathbf{x_i}=\Sigma_j\, a_{ij}\, \mathbf{z_j} \Rightarrow \mathbf{x_i}^2=\Sigma_j\,\Sigma_k\, a_{ij}\; a_{ik}\, \mathbf{z_j}\, \mathbf{z_k}$ Þ

$$E(\mathbf{x_i}^2)=\Sigma_j\,\Sigma_k\, a_{ij}\; a_{ik}\, E(\mathbf{z_j}\,\mathbf{z_k}) \qquad (2.1.5)$$

Pero como las variables $\mathbf{x}_i$ se encuentran tipificadas, y las CPs son incorreladas

$$Var(\mathbf{x_i})=E(\mathbf{x_i}^2),\ E(z_j\, z_k) = \delta_{kj} \qquad (2.1.6)$$

Donde δ_{kj} es la delta de Kronecker.

Por (2.1.5) y (2.1.6), se tiene que

$$Var(\mathbf{x_i})= \Sigma_j\; a_{ij}^{\,2} \qquad (2.1.7)$$

Se tiene que $a_{i1}^{\,2}$ es la contribución de la variable $\mathbf{z_1}$ a la varianza de $\mathbf{x_i}$, y que la suma de las contribuciones de $\mathbf{z_1}$ a la varianza de todas las $\mathbf{x_i}$, viene dada por

$$S= \Sigma_i\; a_{i1}^{\,2} \qquad (2.1.8)$$

Hotelling plantea el problema

$$Max\ \{\Sigma_i\; a_{i1}^{\,2}, \text{ s.a: } r_{ik}= \Sigma_j\, a_{ij}\, a_{kj}\ \} \qquad (2.1.9)$$

Lo resuelve por los multiplicadores de Lagrange y llega a que ese máximo es igual a la mayor de las raíces de la ecuación característica asociada a la matriz de correlaciones.

Las siguientes componentes principales, dice Hotelling, se buscan imponiendo que vayan siendo máximas sus contribuciones a la parte residual de la varianza total pendiente de abarcar.

Teniendo presente que la varianza de una variable, en este caso una CP, es proporcional a la longitud del vector que la representa. El procedimiento anterior es equivalente a este otro proceso iterativo.

Paso 1.- Se considera el conjunto Ω que contiene todos los vectores formados por combinaciones lineales de los vectores $\mathbf{x_j}$. Para encontrar la primera componente principal $\mathbf{z_1}$, cogemos el vector de mayor longitud de Ω; si hay varios candidatos , cogemos cualquiera de ellos; es decir, buscamos una función lineal $\mathbf{a}_1{}^t\mathbf{x}$ de los elementos de $\mathbf{x}$ que tengan máxima varianza, donde $\mathbf{a}_1$ es un vector de *p* constantes, $\alpha_{11}, \alpha_{21}, \ldots, \alpha_{p1}$, de forma que

$$\mathbf{a}_1{}^t\mathbf{x} = \alpha_{11}\mathbf{x}_1 + \alpha_{21}\mathbf{x}_2 + \cdots\cdots + \alpha_{p1}\mathbf{x}_p = \sum_{j=1}^{p} \alpha_{j1}\mathbf{x}_j \ .$$

$\mathbf{a}_1$ ha de ser tal que se maximice la *Var* $(\mathbf{a}_1{}^t\mathbf{x})= \mathbf{a}_1{}^t\mathbf{S}\mathbf{a}_1$. Está claro que, en principio, el máximo no se logrará para $\mathbf{a}_1$ finito, por lo que debe imponerse una restricción de normalización. La restricción que suele imponerse es $\mathbf{a}_1{}^t\ \mathbf{a}_1=1$; es decir, la suma de cuadrados de los elementos de $\mathbf{a}_1$ igualada a 1.

Para maximizar $\mathbf{a}_1{}^t\mathbf{S}\mathbf{a}_1$ sujeto a $\mathbf{a}_1{}^t\mathbf{a}_1=1$, se usa la técnica de los multiplicadores de Lagrange. La función lagrangiana asociada a ese problema es

$$\mathbf{L}(\mathbf{a}_1;\lambda)= \mathbf{a}_1{}^t\mathbf{S}\mathbf{a}_1 - \lambda\,(\mathbf{a}_1{}^t\mathbf{a}_1-1),$$

donde λ es un multiplicador de Lagrange. Diferenciando con respecto a $\mathbf{a}_1$, tenemos

$$2\mathbf{S}\mathbf{a}_1 - 2\,\lambda\,\mathbf{a}_1=0; \text{ es decir, } \mathbf{S}\mathbf{a}_1=\lambda\,\mathbf{a}_1$$

o, dicho de otra forma:

$$(\mathbf{S}-\lambda\mathbf{I_p})\,\mathbf{a}_1=0,$$

donde $\mathbf{I}_p$ es la matriz identidad $(p\times p)$.

Así, λ es un autovalor de **S,** y $\mathbf{a}_1$ es el autovector correspondiente. Obsérvese que la cantidad que se maximiza es

$$\mathbf{a}_1^t \mathbf{S} \mathbf{a}_1 = \mathbf{a}_1^t \lambda \mathbf{a}_1 = \lambda \mathbf{a}_1^t \mathbf{a}_1 = \lambda,$$

y por tanto λ debe ser tan grande como sea posible. De esta manera, $\mathbf{a}_1$ es el autovector que corresponde al autovalor mayor de **S**, y $Var(\mathbf{a}_1^t \mathbf{x}) = \mathbf{a}_1^t \mathbf{S} \mathbf{a}_1 = \lambda_1$ es el mayor autovalor. Llamemos al autovector $\mathbf{v}_1$. Así, la 1ª CP viene dada por $\mathbf{z}_1 = \mathbf{v}_1^t \mathbf{x}$, y su varianza es $var(\mathbf{z}_1) = \lambda_1$.

Una demostración alternativa a la anterior es la siguiente.

S es matriz simétrica, definida positiva; por consiguiente, sabemos que existe una matriz **V** ortonormal, formada por los autovectores de **S**, de forma que $\mathbf{S} = \mathbf{V} \mathbf{L} \mathbf{V}^t$, **L** la matriz de los autovalores de **S**, con lo que

$$\mathbf{V}^t \mathbf{S} \mathbf{V} = \mathbf{L} \qquad (2.1.10).$$

Tomemos el vector $\mathbf{y} = \mathbf{V}^t \mathbf{a}$, con lo que

$$\mathbf{a} = \mathbf{V}\mathbf{y}, \ \mathbf{a}^t = \mathbf{y}^t \mathbf{V}^t, \qquad (2.1.11)$$

y por tanto

$$\mathbf{a}^t \mathbf{S} \mathbf{a} = \mathbf{y}^t \mathbf{V}^t \mathbf{S} \mathbf{V} \mathbf{y} = \text{Por (2.1.10)} = \mathbf{y}^t \mathbf{L} \mathbf{y} = \sum_{i=1}^{p} l_i y_i^2 \quad \text{Þ}$$

$$\mathbf{a}^t \mathbf{S} \mathbf{a} = \sum_{i=1}^{p} l_i y_i^2 \qquad (2.1.12)$$

Por otro lado, $1 = \mathbf{a}^t \mathbf{a} = \text{Por (2.1.11)} = \mathbf{y}^t \mathbf{V}^t \mathbf{V} \mathbf{y} = \mathbf{y}^t \mathbf{y} = \sum_{i=1}^{p} y_i^2$ Þ

$$1 = \mathbf{a}^t \mathbf{a} = \sum_{i=1}^{p} y_i^2 \qquad (2.1.13)$$

Con esto se tiene que nuestro problema de optimización pasa de ser, por (2.1.12) y (2.1.13),

$$Máx\{\mathbf{a}^t \mathbf{S} \mathbf{a}, \text{ s.a.: } 1 = \mathbf{a}^t \mathbf{a} \} \qquad (2.1.14)$$

a ser este otro

$$Máx\{\sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \text{ , s.a.: } 1=\sum_{i=1}^{p} y_i^2 \} \qquad (2.1.15)$$

Supongamos ahora que los autovalores están ordenados decrecientemente. Sea pues λ_1 el mayor autovalor de S, tendremos entonces que

$$\lambda_1 \geq \lambda_i, \forall i \Rightarrow \lambda_1 - \lambda_i \geq 0 \text{ , } \forall I \Rightarrow \sum_{i=1}^{p} (\boldsymbol{l}_i - \boldsymbol{l}_i) y_i^2 \geq 0 \Rightarrow$$

$$\sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \geq \sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \Rightarrow \lambda_1 \sum_{i=1}^{p} y_i^2 \geq \sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \Rightarrow (\text{Por 14}) \Rightarrow$$

$$\lambda_1 \geq \sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \qquad (2.1.16)$$

Pero esa cota superior para $\sum_{i=1}^{p} \boldsymbol{l}_i y_i^2$ se alcanza para $y_1=0$, $y_i=0$, $\forall i \geq 2$; y por tanto se obtiene un máximo para ese punto; es decir para

$\mathbf{a} = \mathbf{V}\mathbf{y} = \mathbf{V}\,(1,0,0,..,0)^t = (\mathbf{v_1},..,\mathbf{v_p})\,(1,0,0,..,0)^t = \mathbf{v_1} \Rightarrow$

$\mathbf{a} = \mathbf{v_1}$ y $Máx\{\mathbf{a}^t\mathbf{S}\mathbf{a}\text{, s.a.: } 1=\mathbf{a}^t\mathbf{a}\} = Máx\{\sum_{i=1}^{p} \boldsymbol{l}_i y_i^2 \text{ , s.a.: } 1=\sum_{i=1}^{p} y_i^2\} = \lambda_1$

Con lo que hemos concluido la demostración alternativa.

Paso 2.- Ahora nos fijamos en el complemento ortogonal de $\mathbf{z_1}$, o sea , en los vectores de Ω que son ortogonales a $\mathbf{z_1}$. El de mayor longitud de estos es el segundo vector componente principal $\mathbf{z_2}$.

Es decir, se busca una función lineal $\mathbf{a}_2^t\mathbf{x}$ incorrelada con $\mathbf{a}_1^t\mathbf{x}$, que tenga máxima varianza, y cumpliendo la restricción de normalización $\mathbf{a}_2^t\mathbf{a}_2 = 1$. Por tanto, la segunda CP, $\mathbf{a}_2^t\mathbf{x}$, maximiza $\mathbf{a}_2^t\mathbf{S}\mathbf{a}_2$ sujeta a ser incorrelada con $\mathbf{a}_1^t\mathbf{x}$ (es decir, sujeta a que la $Cov(\mathbf{a}_1^t\mathbf{x}, \mathbf{a}_2^t\mathbf{x}) = 0$, donde $Cov(\mathbf{x},\mathbf{y})$ denota la covarianza entre las variables $\mathbf{x}$ e $\mathbf{y}$). Pero

$$Cov(\mathbf{a}_1^t\mathbf{x}, \mathbf{a}_2^t\mathbf{x}) = \mathbf{a}_1^t\mathbf{S}\mathbf{a}_2 = \mathbf{a}_2^t\mathbf{S}\mathbf{a}_1 = \mathbf{a}_2^t \lambda_1 \mathbf{a}_1 = \lambda_1 \mathbf{a}_2^t\mathbf{a}_1 = \lambda_1 \mathbf{a}_1^t\mathbf{a}_2$$

Por tanto, las ecuaciones

$$\mathbf{a}_1^t\mathbf{S}\mathbf{a}_2 = 0,\ \mathbf{a}_2^t\mathbf{S}\mathbf{a}_1 = 0,\ \mathbf{a}_2^t\mathbf{a}_1 = 0,\ \mathbf{a}_1^t\mathbf{a}_2 = 0$$

son equivalentes para indicar la incorrelación entre $\mathbf{a}_1^t\mathbf{x}$ y $\mathbf{a}_2^t\mathbf{x}$. el problema de optimización sería:

Hallar $\mathbf{a}_2$, de manera que Máx $\mathbf{a}_2^t\mathbf{S}\mathbf{a}_2$, sujeto a las restricciones $\mathbf{a}_2^t\mathbf{a}_1=0$, $\mathbf{a}_2^t\mathbf{a}_2=1$.

En consecuencia, la función lagrangiana de ese problema es

$$\mathbf{L}(\mathbf{a}_2;\lambda,\phi)= \mathbf{a}_2^t\mathbf{S}\mathbf{a}_2 - \lambda(\mathbf{a}_2^t\mathbf{a}_2-1)-\phi\,\mathbf{a}_2^t\mathbf{a}_1 ,$$

Resolviéndolo, se obtiene que la 2ª CPP viene dada por $\mathbf{z}_2=\mathbf{v}_2^t\mathbf{x}$, y su varianza es var($\mathbf{z}_2$)= $\mathbf{v}_2^t\mathbf{S}\mathbf{v}_2=\lambda_2$, el segundo autovalor más grande de **S**.

Paso 3.- Para la tercera componente principal, limitamos la búsqueda a vectores de Ω que son ortogonales simultáneamente a $\mathbf{z}_1$ y $\mathbf{z}_2$; es decir, ortogonales a la variedad lineal $\mathbf{L}(\mathbf{z}_1 , \mathbf{z}_2)$. El más largo es $\mathbf{z}_3$.

Paso k.- Así vamos procediendo sucesivamente, de manera que en el *k*-ésimo paso, limitamos la búsqueda a vectores de Ω que son ortogonales a $\mathbf{z}_1$,...., $\mathbf{z}_{k-1}$;es decir a L($\mathbf{z}_1$,...., $\mathbf{z}_{k-1}$). El mas largo es $\mathbf{z}_k$.

En consecuencia, se obtiene una función lineal $\mathbf{a}_k^t\mathbf{x}$ que tiene máxima varianza sujeta a ser incorrelada con $\mathbf{a}_1^t\mathbf{x},.....,\mathbf{a}_{k-1}^t\mathbf{x}$, y verificando como siempre la restricción de normalización. En general, la *k*-ésima CP de **x** es $\mathbf{a}_k^t\mathbf{x}$ y *Var(*$\mathbf{a}_k^t\mathbf{x}$ **)**=λ_k , donde λ_k es el *k*-ésimo autovalor más grande de **S** y $\mathbf{a}_k$ es el correspondiente autovector, que llamaremos $\mathbf{v}_k$. Así, la *k*-ésima CP es $\mathbf{z}_k = \mathbf{v}_k^t\mathbf{x}$, y su varianza viene dada por var($\mathbf{z}_k$)= λ_k.

El proceso continúa hasta que Ω queda completamente abarcado, cosa que ocurre, pues el espacio de vectores candidatos se reduce una dimensión en cada paso, con lo que el número de vectores que pueden ser extraídos es igual a la dimensión del espacio V_x generado por los $\mathbf{x}_j$.

3 ACP desde el enfoque de Pearson.-

Como indica Pearson [2], "en muchas investigaciones es deseable representar un sistema de puntos por la recta o plano de mejor ajuste. En casi todos los casos tratados en los libros de textos de mínimos cuadrados, unas variables son tratadas como independientes y otras como dependientes. El resultado es que tenemos una línea recta o plano si tratamos unas

variables como independientes y otra totalmente diferente si tratamos otras variables como las independientes; pues el valor mas probable de *y* para un valor dado de *x* se sabe que no esta dado por la misma relación que el valor más probable de *x* para un valor dado de *y*; es decir, las curvas de regresión empírica de *y* sobre *x* y de *x* sobre *y* no coinciden. En tales casos, los valores de las variables independientes suponemos que son conocidos exactamente, y el valor probable de la variable dependiente es averiguado".

En muchos casos, continua señalando Pearson, "la variable independiente está sujeta a tantas desviaciones o errores como la variable dependiente. Así, nuestra problemática no consiste en que conocemos *x* con precisión y entonces procedemos a encontrar *y*, sino que ambas *x* e *y* son encontradas por experimento u observación. Observamos *x* e *y*, y buscamos una única relación funcional entre ellas. Buscamos una recta o plano de mejor ajuste al sistema de puntos que tengamos". Aunque el termino mejor ajuste es arbitrario, Pearson indica que un buen ajuste puede obtenerse claramente imponiendo que la suma de los cuadrados de las distancias perpendiculares desde los puntos a la recta o plano se haga mínima.

El procedimiento usado por Pearson se diferencia del que seguiremos nosotros en que el problema de optimización que se plantea, lo resolvía directamente por los multiplicadores de Lagrange, mientras que nosotros lo resolveremos reduciendo ese problema a otro equivalente que se resuelve mediante la aproximación de una matriz por otra de igual tamaño pero de menor rango; lo que conduce, en última instancia, al cálculo de la Descomposición en Valores Singulares (DVS) de la matriz de datos.

Sean $P_1,..,P_n$ un sistema de n puntos en un espacio *q* dimensional generado por las variables $\mathbf{x_1,....,x_q}$. Consideremos que las coordenadas del punto P_i son $(x_{i1},..,x_{iq})$. Pearson pretendía encontrar el hiperplano *q*-1 dimensional que más cerca está de esos puntos, en el sentido que la suma de los cuadrados de las distancias perpendiculares de esos puntos al hiperplano fuese mínima. Para ello, tomemos $l_1,...,l_q$ los cosenos directores de un plano que está a una distancia perpendicular d del origen. La ecuación de dicho plano es $l_1\,\mathbf{x_1}+..+\,l_q\mathbf{x_q}=d$, con $l_1^2+....+\,l_q^2=1$.

Por tanto, el problema a resolver pasa de ser buscar el hiperplano Π que verifique

$$\text{Min } \{\textstyle\sum_i d^2(P_i, \Pi)\ \} \qquad (3.1)$$

A este otro

$$\text{Min } \{U=\sum_i(l_1 x_{i1}+..+l_q x_{iq} -d)^2, \text{ s.a.: } l_1^2+..+ l_q^2=1\} \qquad (3.2)$$

Problema que Pearson resuelve directamente usando los multiplicadores de Lagrange, y llegando a la conclusión que si definimos ξ^2 como la media de los cuadrados de los residuales; es decir, ξ^2= U/n, se ha de verificar el siguiente sistema de *q* ecuaciones

$$l_1\sigma_1\,\sigma_i\, r_{1i}+ l_2\sigma_2\,\sigma_i\, r_{2i}+\ldots.+l_i\,(\sigma_i^2-\xi^2)+....+ l_q\sigma_q\,\sigma_i\, r_{qi}=0,\ i=1,..,q \qquad (3.3)$$

Por lo que el mínimo se alcanza cuando ξ^2 es la menor solución de la ecuación:

$$\begin{vmatrix} 1-\boldsymbol{x}^2/\boldsymbol{s}_1^2 & r_{21} & \cdots & r_{q1} \\ r_{12} & 1-\boldsymbol{x}^2/\boldsymbol{s}_2^2 & \cdots & r_{q2} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1q} & r_{2q} & \cdots & 1-\boldsymbol{x}^2/\boldsymbol{s}_q^2 \end{vmatrix} = 0$$

ecuación que se obtiene al imponer que el sistema homogéneo (3.3), que aparece durante el procedimiento, tenga solución distinta de la trivial, con lo que su determinante ha de ser cero.

3.1 Descomposición de una matriz según sus valore s singulares(DVS).-

En 1873, E. Beltrami esbozo el teorema de la D.V.S. para una matriz cuadrada **A**; pero no es hasta 1939 en que en un trabajo de C. Eckart y G. Young aparece el teorema en la forma que ahora es conocido.

Dada la matriz real $\mathbf{X}\in \mathbf{M}_{nxp}$, rango(**X**)=r, sabemos que existen dos matrices ortogonales **U** y **V** tales que

$$\mathbf{X}=\mathbf{U}\begin{bmatrix}\mathbf{D}_r & \theta \\ \theta & \theta\end{bmatrix}\mathbf{V}^t \qquad (3.1.1)$$

donde $\mathbf{D_r}=\text{diag}(\sigma_1,\ldots.,\sigma_r),\ \sigma_1\geq\ldots.\geq\sigma_r>0$;

y donde, expresándolas por columnas, $\mathbf{U}=[\mathbf{u}_1,.....,\mathbf{u}_n]\in \mathbf{M}_{nxn}$ y $\mathbf{V}=[\mathbf{v}_1,.....,\mathbf{v}_p]\in \mathbf{M}_{pxp}$

Los números σ_i, i=1,...,r, ordenados de mayor a menor, y todos mayores que cero, son las raíces cuadradas positivas de los autovalores no nulos de la matriz $\mathbf{X^tX}$, y se denominan valores singulares[1] de la matriz **X**.

La expresión (3.1.1), puede ponerse en la forma

$$\mathbf{X=(U_r,U_{n-r})}\begin{bmatrix}\mathbf{D_r} & \mathbf{q}\\ \mathbf{q} & \mathbf{q}\end{bmatrix}\begin{pmatrix}\mathbf{V_r^t}\\ \mathbf{V_{p-r}^t}\end{pmatrix},$$

donde $\mathbf{U_r}=(\mathbf{u_1},....,\mathbf{u_r})$, $\mathbf{U_{n-r}}=(\mathbf{u_{r+1}},....,\mathbf{u_n})$, $\mathbf{V_r}=(\mathbf{v_1},....,\mathbf{v_r})$, $\mathbf{V_{p-r}}=(\mathbf{v_{r+1}},....,\mathbf{v_P})$. Es decir, que tenemos que

$$\mathbf{X=U_rD_rV_r^t} \qquad (3.1.2),$$

donde $\mathbf{U_r}\in \mathbf{M}_{nxr}, \mathbf{V_r}\in \mathbf{M}_{pxr}, \mathbf{D_r}=diag(\sigma_1,\ldots,\sigma_r)\in \mathbf{M}_{rxr}$, y donde $\mathbf{U_r}$ y $\mathbf{V_r}$, son ortogonales; esto es, $\mathbf{U_r^t\,U_r=I_r}$, $\mathbf{V_r^t\,V_r=I_r}$.

Veamos a continuación algunas propiedades que se derivan de las expresiones anteriores.

1) Las columnas de $\mathbf{U_r}$, $(\mathbf{u_1},....,\mathbf{u_r})$, constituyen una base ortonormal de Img(**X**), es decir, del espacio engendrado por los vectores columnas de **X**.

   En efecto, cualquier vector perteneciente al espacio de las columnas de **X**, $\mathbf{y\,\hat{I}\,\hat{A}^p}$, es de la forma:

$$\mathbf{y}=\sum_{j=1}^{p} a_j\mathbf{x_j}=\mathbf{Xa}=\text{Por (3.1.2)}=\mathbf{U_rD_rV_r^ta}=\sum_{i=1}^{r} s_i(\mathbf{v_i^ta})\mathbf{u}_i$$

2) Las columnas de $\mathbf{V_{p-r}}=(\mathbf{v_{r+1}},....,\mathbf{v_p})$, forman una base ortonormal de *Ker*(**X**)=$\{\mathbf{y\,\hat{I}\,\hat{A}^p}$ /**X**.**y**=**0**$\}$; es decir, del espacio de los vectores ortogonales al espacio engendrado por las filas de **X**.

   En efecto,

   $\mathbf{X.y=0}$ **,** y por tanto,

[1] En algunas ocasiones representaremos σ^2/n como λ; o lo que es igual $\sigma/n^{1/2}=\lambda^{1/2}$.

$$\mathbf{X}\cdot\mathbf{y}=\mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^t\mathbf{y}=\sum_{i=1}^{r}\sigma_i(\mathbf{v}_i^t\mathbf{y})\mathbf{u}_i=\mathbf{0}\Leftrightarrow$$

$$\mathbf{v}_i^t\mathbf{y}=0,\forall i=1,..,r\Leftrightarrow \mathbf{y}\perp\mathbf{v}_i,\forall i=1,..,r$$

3) Las columnas de $\mathbf{U_{n-r}}=(\mathbf{u_{r+1}},....,\mathbf{u_n})$ forman una base ortonormal de Ker($\mathbf{X^t}$)={$\mathbf{y}\in\mathbb{R}^n$ /$\mathbf{X^t.y=0}$}; es decir del espacio ortogonal al espacio de las columnas de **X**.

4) Las columnas de $\mathbf{V_r}=(\mathbf{v_1},....,\mathbf{v_r})$ forman una base ortonormal de Img($\mathbf{X^t}$); es decir del espacio engendrado por las filas de **X**.

5) $\mathbf{X=U_rD_rV_r^t}=(\mathbf{u_1},....,\mathbf{u_r})\begin{pmatrix}\sigma_1 & & 0\\ & \ddots & \\ 0 & & \sigma_r\end{pmatrix}\begin{pmatrix}\mathbf{v_1^t}\\ \vdots\\ \mathbf{v_r^t}\end{pmatrix}=\sum_{i=1}^{r}\sigma_i\mathbf{u_i}\mathbf{v_i^t}$; es decir,

X puede expresarse como suma de r matrices de rango uno. Pues por ejemplo,

$$\mathbf{u_1}\cdot\mathbf{v_1}^t=\begin{pmatrix}u_{11}\\ u_{12}\\ \vdots\\ u_{1n}\end{pmatrix}\begin{pmatrix}v_{11} & v_{12} & \cdots & v_{1p}\end{pmatrix}=\begin{pmatrix}u_{11}v_{11} & u_{11}v_{12} & & u_{11}v_{1p}\\ u_{12}v_{11} & u_{12}v_{12} & & u_{12}v_{1p}\\ & & & \\ u_{1n}v_{11} & u_{1n}v_{12} & & u_{1n}v_{1p}\end{pmatrix}$$, que es de

rango uno pues todas las columnas son proporcionales a la primera.

6) Las columnas de $\mathbf{V_r}=(\mathbf{v_1},....,\mathbf{v_r})$ son los autovectores de $\mathbf{X^tX}$, con autovalores asociados no nulos $\sigma_1^2,\ldots,\sigma_r^2$, respectivamente, pues, por (3.1.2), $\mathbf{X^tX=V_rD_r^2V_r^t}$, y por tanto $\mathbf{(X^tX)V_r=V_rD_r^2}$, de donde $\mathbf{(X^tX)v_i=v_i}\sigma_i^2$, *i*=1,..,r.

7) Las columnas de $\mathbf{U_r}=(\mathbf{u_1},....,\mathbf{u_r})$ son los autovectores de $\mathbf{XX^t}$, con autovalores asociados no nulos $\sigma_1^2,\ldots,\sigma_r^2$, respectivamente, pues, por (3.1.2); $\mathbf{XX^t=U_rD_r^2U_r^t}$., y por tanto $\mathbf{(XX^t)U_r=U_rD_r^2}$, de donde $\mathbf{(XX^t)u_i=u_i}\sigma_i^2$, *i*=1,..,r.

8) Se verifica que

$$\|\mathbf{X}\|^2 = Traza(\mathbf{X}^t\mathbf{X}) = Traza(\mathbf{V_r}\mathbf{D_r^2}\mathbf{V_r^t}) =^{2} Traza(\mathbf{D_r^2}) = \sum_{i=1}^{r} \boldsymbol{s}_i^{\,2}$$

3.2 Aproximación matricial a una matriz X por otra de rango menor.-

El antecedente histórico de este problema fue introducido por primera vez , en 1907, en un trabajo sobre ecuaciones integrales lineales; posteriormente, en 1936-1938 C.Eckart [3], G.Young y A.S. Householder [4], plantearon el problema siguiente:

Dado $\mathbf{X}\,\hat{\mathbf{I}}\,\hat{\mathbf{A}}^{\mathbf{nxp}}$ $rg(\mathbf{X})$=r, se trata de hallar

$$\underset{\mathbf{X}^* \in \Re^{nxp},\, rg\,(\mathbf{X}^*)=k<r}{Min} \|\mathbf{X}-\mathbf{X}^*\| \qquad (3.2.1)$$

donde $\mathbf{X}\,\hat{\mathbf{I}}\,\hat{\mathbf{A}}^{\mathbf{nxp}}$ es fija, y la $\|\mathbf{X}\|$ es la norma de Frobenius; es decir $\|\mathbf{X}\| = \sqrt{\sum_{i,j} x_{ij}^2}$.

La solución viene dada en términos de los valores singulares de **X**.

Puede probarse que

$$\underset{\mathbf{X}^* \in \Re^{nxp},\, rg\,(\mathbf{X}^*)=k<r}{Min} \|\mathbf{X}-\mathbf{X}^*\| = \|\mathbf{X}-{}_{\mathbf{k}}\mathbf{X}\| = (\boldsymbol{s}_{k+1}{}^2 + .. + \boldsymbol{s}_r^{\,2})^{1/2} \qquad (3.2.2)$$

donde,

$${}_{\mathbf{k}}\mathbf{X} = \mathbf{U_k}\mathbf{D_k}\mathbf{V_k^t} = \sum_{i=1}^{k} \boldsymbol{s}_i \mathbf{u_i}\mathbf{v_i^t} = \mathbf{U}\mathbf{D_k}\mathbf{V^t} \qquad (3.2.3)$$

y donde se tiene que $\mathbf{D_k} = \begin{pmatrix} \boldsymbol{s}_1 & & & q \\ & \ddots & & \\ & & \boldsymbol{s}_k & \\ q & & & q \end{pmatrix}$.

${}_{\mathbf{k}}\mathbf{X}$ se llama aproximación matricial de rango k de la matriz **X**; y la matriz $\mathbf{X}-{}_{\mathbf{k}}\mathbf{X}$ se llama matriz de residuos.

[2] Harville, 1999

Obsérvese que mientras que las filas de $\mathbf{X}$ son una nube de puntos en un espacio euclídeo p-dimensional, las filas de $_k\mathbf{X}$ son puntos desconocidos en un espacio k-dimensional. $_k\mathbf{X}=\mathbf{U_k D_k V_k}$ identifica el subespacio que está más próximo a la nube de puntos.

Los vectores $\{\mathbf{v_1,.....,v_k}\}$, columnas de $\mathbf{V_k}$, definen los ejes principales ortonormales del subespacio $L(\mathbf{v_1,....,v_k})$.

Las filas de la matriz $\mathbf{U_k D_k}=[\sigma_1\mathbf{u_1},\ldots\ldots,\sigma_k\mathbf{u_k}]$ definen las coordenadas, respecto a esos ejes, de las proyecciones de la nube de puntos sobre el subespacio.

Tenemos que conseguir que los subespacios origen y proyección estén lo más próximo posible; es decir, que $\|\mathbf{X}-{}_k\mathbf{X}\|$ sea lo más pequeña posible.

La bondad de la aproximación viene dada por el cociente entre la varianza explicada por la aproximación y la varianza total; es decir, por el cociente entre la norma de la matriz $_k\mathbf{X}$ y la norma de la matriz $\mathbf{X}$.

$$\frac{\text{VE}}{\text{VT}}=\frac{\|{}_k\mathbf{X}\|^2}{\|\mathbf{X}\|}=\frac{Traza({}_k\mathbf{X}^t\,{}_k\mathbf{X})}{Traza(\mathbf{X}^t\mathbf{X})}=\frac{\sum_{i=1}^{k}s_i^2}{\sum_{i=1}^{r}s_i^2} \qquad (3.2.4)$$

También puede expresarse la bondad de la aproximación como:

$$1-\frac{\text{VNE}}{\text{VT}}=1-\frac{\|\mathbf{X}-{}_k\mathbf{X}\|}{\|\mathbf{X}\|}=\cdots\cdots=1-\frac{\sum_{i=k+1}^{r}s_i^2}{\sum_{i=1}^{r}s_i^2} \qquad (3.2.5)$$

3.3 Teorema de construcción de las CPs.-

Supongamos que las observaciones iniciales $\mathbf{f_1,..,f_n}$ *se transforman por* $\mathbf{f_i^*=B^t f_i}$, i=1,2..,n, *donde* $\mathbf{B}$ *es una matriz pxq con columnas ortonormales, así que* $\mathbf{f_1^*,..,f_n^*}$ *son las proyecciones ortogonales de* $\mathbf{f_1,..,f_n}$ *sobre el subespacio q-dimensional generado por las columnas de la matriz* $\mathbf{B}$. *Si la 'bondad del ajuste' de este subespacio q-dimensional a los puntos* $\mathbf{f_1,..,f_n}$ *se define como la suma de los cuadrados de las distancias perpendiculares de los puntos* $\mathbf{f_1,..,f_n}$ *al subespacio, se verifica que esta medida se minimiza cuando* $\mathbf{B}=\mathbf{V_q}$.

Demostración.-

Notemos por L(**B**) el subespacio generado por las columnas de **B**.

La nueva matriz de datos es $\mathbf{X}^*=\mathbf{XB}$, datos cuyas coordenadas están dadas en la nueva base.

Asi, los $\mathbf{f_i}^*$ son las coordenas de los puntos $\mathbf{f_i}$ en la nueva base $\{\mathbf{b_1},...,\mathbf{b_q}\}$, donde $\mathbf{b_j}$ es la columna j-ésima de la matriz **B**.

Queremos minimizar $\sum_{i=1}^{n} d^2(\mathbf{f}_i, L(\mathbf{B}))$

Sea $\mathbf{f_i}^*=Proy_{L(\mathbf{B})}\mathbf{f_i}$, en coordenas de la base formada por las columnas de la matriz **B**.

Sea $\mathbf{X}^*$ la matriz cuyas filas son los $\mathbf{f_i}^{*t}$. En consecuencia, $\mathbf{X}^*=\mathbf{XB}$ y rango($\mathbf{X}^*$)=rango(**B**)=q.

Por consiguiente, tenemos que

$d^2(\mathbf{f_i}, L(\mathbf{B}))= d^2(\mathbf{f_i}, Proy_{L(\mathbf{B})}\mathbf{f_i})= d^2(\mathbf{f_i}, \mathbf{f_i}^*)=\left\|\mathbf{f}_i - \mathbf{f}_i^*\right\|^2$; luego,

$\sum_{i=1}^{n} d^2(\mathbf{f_i}, L(\mathbf{B})) = \sum_{i=1}^{n}\left\|\mathbf{f_i} - \mathbf{f_i}^*\right\|^2 = \left\|\mathbf{X}-\mathbf{X}^*\right\|^2$. En virtud de ese resultado y dado que queríamos minimizar $\sum_{i=1}^{n} d^2(\mathbf{f}_i, L(\mathbf{B}))$, lo que tenemos que hacer es hallar la matriz $\mathbf{X}^* \in \hat{\mathbf{A}}^{nxp}$, rg($\mathbf{X}^*$)=q≤p, que haga mínimo $\left\|\mathbf{X}-\mathbf{X}^*\right\|^2$.

La solución a ese problema viene dada mediante la Descomposición de una matriz según sus Valores Singulares, y en concreto mediante la aproximación matricial de rango q de la matriz **X**. Así, se tiene que la solución es

$$\mathbf{X}^*=\mathbf{U_q D_q V_q}^t=\begin{pmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_q\end{pmatrix}\begin{pmatrix} s_1 & & q \\ & & \\ q & & s_q\end{pmatrix}\begin{pmatrix}\mathbf{v}_1^t \\ \vdots \\ \mathbf{v}_q^t\end{pmatrix}$$

Donde los números σ_i, i=1,..,p son todos mayores que cero, están ordenados de mayor a menor, y son los valores singulares de la matriz **X** (las raíces cuadradas positivas de los autovalores no nulos de la matriz $\mathbf{X^tX}$).

Además, sabemos que las columnas de $\mathbf{V_q}=\{\mathbf{v_1},..,\mathbf{v_q}\}$ forman una base ortonormal para el subespacio de las filas de $\mathbf{X}^*$ y, en consecuencia, $\mathbf{f_i}^*=\mathbf{V_q}^t\mathbf{f_i}$ son las coordenadas de los puntos

$\mathbf{f_i}$ en la nueva base; es decir, son las coordenadas de las proyecciones de los puntos $\mathbf{f_i}$ sobre los nuevos ejes.

Por tanto se concluye que $\mathbf{B}=\mathbf{V_q}$, como queríamos demostrar.#

Además, la matriz de los individuos proyectados es $\mathbf{Z}=\mathbf{XV_q}$, que es una matriz cuyas filas nos dan las coordenadas de los puntos proyectados respecto a la nueva base.

En consecuencia, las nuevas variables se corresponden con las columnas de la nueva matriz de datos $\mathbf{Z}=\mathbf{XV_q}$.

Mediante la Descomposición en Valores Singulares hemos obtenido que $\mathbf{X}=\mathbf{UDV}^t \Rightarrow \mathbf{XV}=\mathbf{UD} \text{ Þ } \mathbf{XV_q}=\mathbf{U_qD_q}$

Y por tanto, se tiene que $\mathbf{Z}=\mathbf{XV_q}=\mathbf{U_qD_q}=(\sigma_1\mathbf{u_1},..,\sigma_q\mathbf{u_q})$, con lo que las nuevas variables son $\mathbf{z_k}=\sigma_k\mathbf{u_k}$, $k=1,..,q$. Además, las filas de esa matriz $\mathbf{U_qD_q}$ **(Z)** definen las coordenadas respecto a la base $\mathbf{V_q}=\{\mathbf{v_1},..,\mathbf{v_q}\}$ de las proyecciones de las filas de $\mathbf{X}$ sobre el subespacio que determina $\mathbf{V_q}$. En consecuencia, tenemos que :

$$\underset{\mathbf{L(v_1,..,v_q)}}{Proy}\ \mathbf{f_i} = (s_1 u_{i1})\mathbf{v}_1 + ... + (s_q u_{iq})\mathbf{v_q}$$

En resumen, las Componentes Principales se definen como las variables obtenidas al proyectar los individuos $\mathbf{f_1},..,\mathbf{f_n}$ sobre los subespacios de dimensión *1,2,..,p-1* más próximos a la nube de puntos que determinan dichos individuos $\mathbf{f_1},..,\mathbf{f_n}$.

El mejor ajuste lineal determina la primera Componente Principal, y la dirección de la última C.P. es la dirección ortogonal al mejor hiperplano de ajuste (p-1)-dimensional.

Los subespacios anteriores, que hemos calificado como "más próximos" son aquellos para los cuales la suma de los cuadrados de las distancias perpendiculares de los individuos $\mathbf{f_1},..,\mathbf{f_n}$ al subespacio se minimiza.

De hecho, éste es, en lo fundamental, el acercamiento adoptado por Pearson (1901), aunque él se concentró en dos casos especiales donde $q = 1$ y $q = p - 1$; es decir, él busco la mejor recta de ajuste y el mejor subespacio de una dimensión menos que el espacio original en el que están inmersos los individuos.

3.4 Equivalencia entre los enfoques de Pearson y de Hotelling.-

Veamos a continuación, que las Componentes Principales obtenidas según el enfoque de Pearson cumplen los principios exigidos por Hotelling, cuando fueron definidas por él:

1) *Las nuevas variables se expresan como combinación lineal de las antiguas y recíprocamente.*

En efecto, las nuevas variables, $\mathbf{z_k}$, son las columnas de la matriz

$\mathbf{Z=XV=X}\,(\mathbf{v_1},..,\mathbf{v_p}) \Rightarrow \mathbf{z_k=X\,v_k}$, $k=1,..,p \Rightarrow$

$$\mathbf{z_k=X\,v_k=(x_1,..,x_p)}\begin{pmatrix} v_{1k} \\ \vdots \\ v_{pk} \end{pmatrix} = \sum_{i=1}^{p} v_{ik}\mathbf{x}_i \quad \text{Þ}$$

$$\mathbf{z}_k = \sum_{i=1}^{p} v_{ik}\mathbf{x}_i\,, \qquad k=1,..,p. \tag{3.4.1}$$

Recíprocamente, veamos las variables originales, $\mathbf{x_j}$, como combinación lineal de las nuevas, $\mathbf{z_k}$.

Sabemos por la D.V.S. que

$$\mathbf{X=UDV^t}= \sum_{k=1}^{p} s_k \mathbf{u_k}\,\mathbf{v_k^t} \tag{3.4.2}$$

Pero como $\mathbf{X=UDV^t}$ Þ $\mathbf{XV=UD}$ Þ $\mathbf{z_k=Xv_k=u_k}\sigma_k$,k=1,..,$p$; luego, por (3.4.2), tendremos que

$$\mathbf{X}= \sum_{k=1}^{p} (\mathbf{Xv_k})\mathbf{v_k^t} = \sum_{k=1}^{p} \mathbf{z_k}\,\mathbf{v_k^t} \quad \text{Þ} \quad \mathbf{X}= \sum_{k=1}^{p} \mathbf{z_k}\,\mathbf{v_k^t} \quad \text{Þ}$$

$$\mathbf{x_j}= \sum_{k=1}^{p} v_{jk}\,\mathbf{z_k}\,, \forall\, j=1,..,p \tag{3.4.3}$$

2) *Las nuevas variables son incorreladas entre sí.*

En efecto, $\mathbf{z_k}=\sigma_k\mathbf{u_k}$, luego $\mathbf{z_k}^t\,\mathbf{z_h}=(\sigma_k\mathbf{u_k}\,)^t(\sigma_h\mathbf{u_h})= \sigma_k\mathbf{u_k}^t\,\mathbf{u_k}\sigma_k=0$, pues los $\mathbf{u_k}$ eran incorrelados. En consecuencia, los $\mathbf{z_k}$ también son incorrelados.

3) *Veamos que las nuevas variables están ordenadas decrecientemente según sus varianzas*.

$\mathbf{z_k}=\sigma_k\mathbf{u_k}\Rightarrow$ $\text{Var}(\mathbf{z_k})=\frac{1}{n}\|\mathbf{z}_\mathbf{k}\|^2=\frac{1}{n}\mathbf{z}_\mathbf{k}{}^t\mathbf{z}_\mathbf{k}=\frac{1}{n}s_k\mathbf{u}_\mathbf{k}{}^t\mathbf{u}_\mathbf{k}s_k=\frac{1}{n}s_k{}^2$; en consecuencia,

$\text{Var}(\mathbf{z_k})=\frac{1}{n}s_k{}^2$, $k=1,..,p$; y por tanto, dado que los σ_k estaban ordenados decrecientemente, las variables $\mathbf{z_k}$ están ordenadas según la magnitud de sus varianzas.

En consecuencia, la nueva matriz de varianzas-covarianzas viene dada por

$$\mathring{\mathbf{a}}_\mathbf{Z}=\frac{1}{n}\begin{pmatrix} s_1^2 & & q \\ & \ddots & \\ q & & s_p^2 \end{pmatrix}=\frac{1}{n}\mathbf{D}_p^2 .$$

La obtención de esta matriz $\mathbf{S_Z}$, también puede hacerse por este otro camino alternativo:

$$\frac{1}{n}(\mathbf{Z}^t\mathbf{Z})=\frac{1}{n}((\mathbf{UD})^t(\mathbf{UD}))=\frac{1}{n}(\mathbf{D}^t\mathbf{U}^t\mathbf{UD})=\frac{1}{n}(\mathbf{D}^t\mathbf{D})=\frac{1}{n}\mathbf{D}^2$$

4) *Veamos seguidamente que la varianza total y generalizada de ambos conjuntos de variables (el inicial y el de componentes principales) no ha variado en el caso de elegir el mismo número de Componentes Principales que de variables originales.*

En efecto, por un lado tenemos que la suma de las varianzas de las nuevas variables es

$\sum_{k=1}^{p}\text{Var}(\mathbf{z}_k)=\sum_{k=1}^{p}\frac{1}{n}s_k{}^2$, por tanto tenemos que

$$\sum_{k=1}^{p}Var(\mathbf{z}_k)=\sum_{k=1}^{p}\frac{1}{n}s_i^2=\frac{1}{n}\sum_{k=1}^{p}s_i^2=\textit{Por la}\ \text{propiedad 8 de la D.V.S.}=$$
$$=\frac{1}{n}traza(\mathbf{X}\mathbf{X}^t)=traza(\frac{1}{n}\mathbf{X}^t\mathbf{X})=traza(\Sigma_\mathbf{X})=\sum_{i=1}^{p}Var(\mathbf{x}_i)\Rightarrow$$
$$\Rightarrow\sum_{k=1}^{p}Var(\mathbf{z}_k)=\sum_{i=1}^{p}Var(\mathbf{x}_i)\ ,\ \text{como queriamos demostrar.}$$

Y, por otro lado, tenemos que

$$det(\acute{\mathbf{O}}_\mathbf{Z})=det(\frac{1}{n}\mathbf{D}^2)=\frac{1}{n^p}\prod_{i=1}^{p}s_i^2=\frac{1}{n^p}det(\mathbf{X}^t\mathbf{X})=det(\frac{1}{n}\mathbf{X}^t\mathbf{X})=det(\acute{\mathbf{O}}_\mathbf{X})$$

de donde $det(\acute{\mathbf{O}}_\mathbf{Z})=det(\acute{\mathbf{O}}_\mathbf{X})$, como también queríamos probar.

5) *Veamos a continuación cual es la bondad de la aproximación realizada al quedarnos solo con las q primeras componentes principales*.

$$\frac{Var.\ actual}{Var.\ total} = \frac{\left\| {}_q\mathbf{X} \right\|^2}{\left\| \mathbf{X} \right\|^2} = \frac{Traza\left(\mathbf{X_q^t X_q} \right)}{Traza\left(\mathbf{X^t X} \right)} = \frac{\sum_{k=1}^{q} s_k^{\,2}}{\sum_{k=1}^{r} s_k^{\,2}} = \frac{\frac{1}{n}\sum_{k=1}^{q} s_k^{\,2}}{\frac{1}{n}\sum_{k=1}^{r} s_k^{\,2}} = \frac{\sum_{k=1}^{q} Var\left(\mathbf{z_k} \right)}{\sum_{k=1}^{r} Var\left(\mathbf{c_k} \right)}$$

Bibliografía.-

[1] Hotelling, 1933. Analysis of a complex of statistical variables into principal components. *Journal educ. Psychol.*,24,417-441, 498-520

[2] Pearson, k, 1901. On lines and planes of closest fit to systems of points in space. *Phil. Maga.*, 2, 559-572.

[3] Eckart,C and Young, G, 1936. Approximation of one matrix by another of lower rank. *Psycometrika*,1, 211-218.

[4] Householder, A.S. and Young, G, 1938. Matrix approximation and latent roots. *Ame. Math. Mon.*, 45, 165-171