Introducción

Cuando afrontamos un Análisis Multivariante de datos, el escenario típico suele estar constituido por una masa de datos generalmente grande no sólo porque suele proceder de muchos individuos (muchos casos) sino también porque sobre cada uno de esos individuos se suelen medir un número sustancial de variables.

Generalmente, la información que proporcionan estas "muchas" variables suele ser en buena parte redundante al presentarse entre ellas múltiples relaciones de dependencia manifestadas por la existencia de correlaciones considerables. Así, explicar el comportamiento de los datos, de una forma clara (o al menos sencilla), a partir de esas variables inicialmente observadas y altamente correlacionadas resulta una tarea dificultosa.

Las técnicas factoriales pretenden, desde sus diferentes enfoques, abordar el problema de simplificar la interpretación del comportamiento observado de los datos.

Para ilustrar brevemente algunos de estos enfoques, imaginemos que disponemos de las calificaciones en nueve asignaturas de los 29 alumnos de un curso, según se indica en el cuadro siguiente (lista detallada de datos en el ejemplo al final del tema):

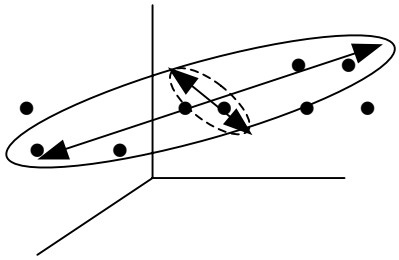
Calificaciones de los alumnos de un curso

Caso	ST1	ST2	GES	ST3	IOP	INF	MAT	ECO	ING
1	,3	,3	1,0	,0	1,7	,6	,6	,6	,3
...	...	...	...	...	...	...	...	...	...
15	6,2	6,5	4,1	4,2	2,4	3,4	5,5	6,2	5,8
...	...	...	...	...	...	...	...	...	...
29	10,0	10,0	9,6	8,7	9,6	9,3	10,0	7,2	7,5

variables: Estadística 1 (ST1), Estadística 2 (ST2), Estadística 3 (ST3), Investigacion Operativa (IOP), Informática (INF), Matemáticas (MAT), Economía (ECO), Gestión (GES) e Inglés (ING).

Ya en una primera aproximación podemos comprobar la dificultad de visualizar esta información de manera completa. Nuestra limitada percepción intuitiva de las cosas, acostumbrada a espacios físicos de 3 dimensiones (o a lo sumo de 4 si incorporamos el tiempo), puede permitirnos imaginar la existencia de un espacio de nueve dimensiones como el de nuestro ejemplo, pero difícilmente nos permite visualizar lo que ocurre en él y que los datos manifiestan.

Podemos tratar de vislumbrar este comportamiento global en ese espacio complejo, a partir de sus proyecciones resultantes sobre los subespacios formados por cada dos ó tres de esas nueve variables; lo que podemos representar y comprender bastante bien mediante gráficos bidimensionales o tridimensionales.



En estas representaciones, las nubes de puntos proyectados aparecerán más alargada en aquella dirección donde se presente una mayor dispersión o variabilidad (en general, mayor variedad o diversidad) de los datos, y menos alargada en aquella dirección donde haya una menor dispersión o variabilidad (en general, menor variedad o diversidad) de los datos, como intuitivamente puede verse en el anterior gráfico.

En cualquier caso, con las 9 variables originales podemos construir 84 proyecciones tridimensionales sustancialmente diferentes. Surge inmediatamente la necesidad de simplificar este enfoque: para empezar, ¿cuál de todas estas proyecciones refleja mejor la realidad global? ¿se pierde mucha información? Y si no nos restringirnos a las variables originales, ¿existen proyecciones más fidedignas sobre otros subespacios? Desde una óptica intuitiva, 3 dimensiones son deseables, pero ¿son suficientes para reflejar la realidad con cierta precisión?

Si nuestro objetivo es llegar a comprender de una forma sencilla y simplificada a qué se debe la diversidad de calificaciones que se observan, intuitivamente podríamos estar de acuerdo en que los alumnos más inteligentes y con mejor predisposición al estudio tendrán generalmente mejores notas en la mayoría de las asignaturas; o en que los alumnos con peor formación cuantitativa probablemente tendrán más problemas a la hora de sacar buenas notas en asignaturas como las matemáticas o las estadísticas. Al hacer este razonamiento estamos implícitamente admitiendo que probablemente existen unas variables (factores, componentes,...), probablemente no observadas directamente, y que, de forma causal o no, permiten "simplificar" la explicación de los comportamientos observados.

Así el Análisis de Componentes Principales simplemente se pregunta por cuántas y cuáles serían esas pocas variables que nos permitirían resumir la diversidad de las calificaciones observadas con la menor pérdida de información posible. Por su parte, el Análisis Factorial presupone la existencia de un número pequeño de variables no observables o latentes (factores) que serían la causa de las calificaciones observadas y que trata de identificar.

Aproximación al Análisis de Componentes Principales

El Análisis de Componentes Principales (en adelante, ACP) tratará de representar, "de forma clara y ordenada", la variedad de los comportamientos observados en un conjunto de n individuos mediante un conjunto de p variables. Es decir, buscará un nuevo sistema de ejes coordenados, ordenados, (nuevas variables de referencia que llamaremos *componentes principales*) con el que poder apreciar y analizar más claramente la diversidad de comportamiento reflejada en los datos. Para ello, determinará como primer eje coordenado la nueva variable (primera componente principal) que explique la máxima variabilidad (diversidad) posible de los datos observados; para proceder secuencialmente y de forma

análoga a determinar los sucesivos ejes coordenados (sucesivas componentes principales) a partir del resto de la variabilidad (diversidad) de los datos, aún no explicada por los anteriores.

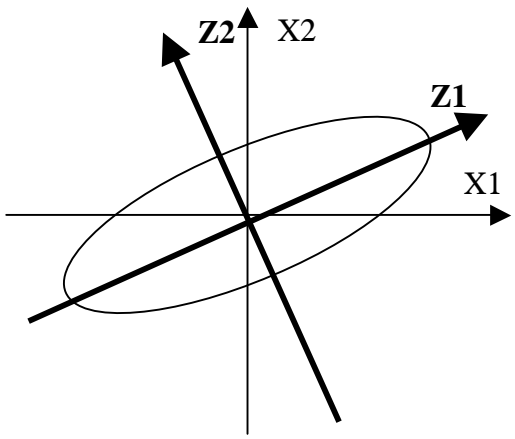
Así, siguiendo con nuestro ejemplo, el ACP tratará de responder a la pregunta ¿en qué sistema de nuevos ejes coordenados podríamos apreciar de una forma más clara y ordenada la diversidad de las calificaciones?

Si representamos por $X_1,...,X_p$ las variables originales y nuestro objetivo es pues, encontrar unas nuevas variables (componentes principales) $Z_1,...,Z_p$ que nos expliquen ordenadamente y de la forma más clara la variabilidad de los datos, parece lógico determinar la primera componente principal Z_1 como aquella que vaya en la dirección de máxima variabilidad de los datos y que, por tanto, explicará la mayor diversidad entre los datos; ya que los datos se dispersan de una forma máxima justamente en esa dirección. Esta dirección, pues, nos informará mucho del comportamiento más diversamente llamativo de esa nube de puntos.

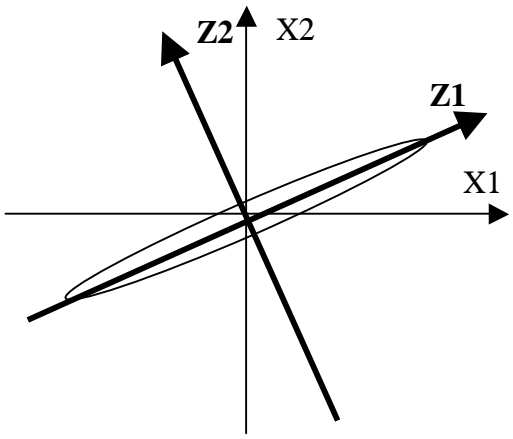
Por otra parte, obsérvese que para que estas nuevas variables de referencia (nuevo sistema de ejes coordenados) permita una representación "clara" de la realidad, deberíamos pedir lógicamente que estuviesen incorrelacionadas para que cada nueva variable informara de aspectos diferentes de la realidad y así facilitar la interpretración. Recordemos que nubes de puntos inclinadas indicaban correlación entre variables y que nubes de puntos paralelas a los ejes indicaban incorrelación entre variables, por lo que la incorrelación entre las nuevas variables de referencia (componentes principales) se conseguirá cuando se tomen paralelas a los ejes principales de la nube de punto. Ello nos induce a pensar que si la nube de puntos es lo suficientemente regular (aproximadamente elipsoidal), la direcciones de las componentes principales deben ser sus ejes ortogonales.

Así pues, la variable Z_2 deberá ser la variable que, siendo ortogonal a Z_1, tenga la dirección de máxima dispersión de las restantes. Así aportará una información adicional del resto de la variabilidad de los datos y que no quedaba explicada por la dirección Z_1 (nótese que existe toda una gama de individuos con un mismo valor para Z_1 que pueden presentar diferentes valores para Z_2).

Intuitivamente, este proceso puede verse reflejado en la siguiente figura.



Secuencialmente, las sucesivas componentes principales irán perdiendo importancia explicativa de la diversidad o variabilidad de los datos, ya que se extienden en direcciones de cada vez menos dispersión. Ello se acentuará más cuanto mayor sea la correlación entre las variables originales ya que cuanta mayor dependencia haya entre ellas, más alargada será la nube de puntos en alguna dirección y más estrecha en alguna dirección perpendicular (suponiendo siempre que la relación entre ellas fuera lineal), como se aprecia en la siguiente figura.



En el caso límite de que esa regresión fuera perfecta, y por tanto todos los puntos estuvieran sobre un hiperplano, la componente principal perpendicular al hiperplano no aportaría ninguna información porque no habría variabilidad en su dirección. Es en estos casos cuando vamos a conseguir una reducción efectiva de la dimensión de nuestro problema, al poder obviar o suprimir las componentes principales que no aportan información sobre la diversidad.

Así que, como consecuencia del proceso, el ACP no sólo encuentra ordenadamente las direcciones que mejor explican la variabilidad de esa nube de puntos, sino que también, en el caso de que haya información redundante, permitirá prescindir de alguna de las últimas componentes, bien porque estrictamente no expliquen nada acerca de la variación de los datos, o bien porque expliquen una cantidad despreciable de la misma, consiguiendo simplificar el problema mediante la reducción efectiva de la dimensión del mismo.

La Técnica del Análisis de Componentes Principales

El ACP es desarrollado por Hotelling en 1933, aunque sus orígenes se remontan a los trabajos pioneros, en 1901, sobre ajustes por mínimos cuadrados ortogonales de Karl Pearson.

Supongamos n indidividuos sobre los que se han observado p variables X_1, X_2,,...,X_p. Sin pérdida de generalidad supongamos que éstas están *centradas*; es decir, tienen marginalmente medias iguales a cero. Y notemos por X a la matriz de datos original de dimensión $n \cdot p$ (centrada, por tanto, en nuestro caso)

La técnica del ACP busca la solución del siguiente problema:

Encontrar unas nuevas variables Z_1, Z_2,,…,Z_p (componentes principales), nuevo sistema coordenado, tales que:

1º.- son combinaciones lineales de las originales $Z_i = X_1u_{1i}+...+X_p u_{pi} = Xu_i$, i=1,...,*p*

- consecuentemente, las nuevas variables Z1, Z2,,...,Zp están igualmente centradas y se cruzan en el origen de coordenadas.
- además, los vectores $u_i = (u_{1i}, ..., u_{pi})'$ son los vectores de dirección de las componentes principales, Z_i, en el espacio de las X_1,...,X_p.

2º.- con $|u_1| = |u_2| = ... = |u_p| = 1$ ($\Leftrightarrow u_i' \cdot u_i = 1$)

- De todos los posibles vectores que marcan una misma dirección, tomaremos los vectores unitarios, los de módulo 1, para evitar su indeterminación.
- Consecuentemente, z_{ij} es la proyecciones del caso i en el nuevo eje Zj

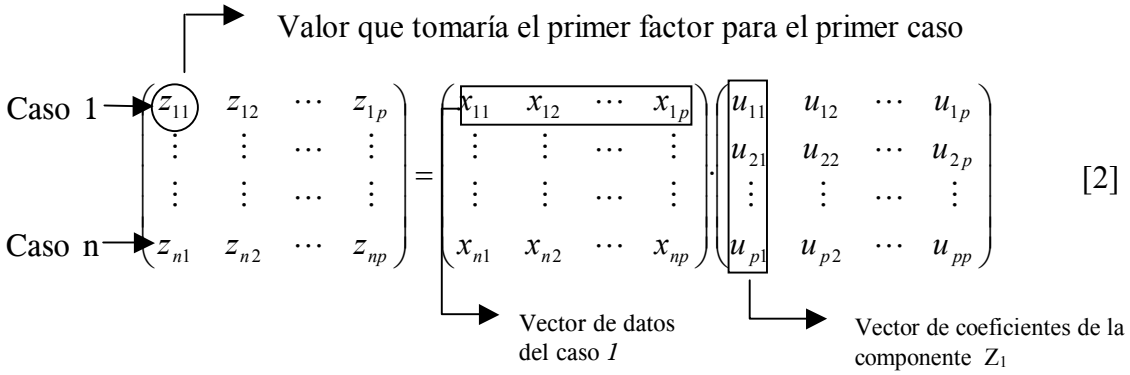
3º.- u_i ortogonal a u_j, para todo $i \neq j$ ($\Leftrightarrow u_i' \cdot u_j = 0$)

- Los nuevos ejes son ortogonales

4º.- $var(Z_1) \geq var(Z_2) \geq ... \geq var(Z_p)$

- Los nuevos ejes se extraen ordenadamente, de más a menos explicativos de la diversidad

Pormenorizadamente, la expresión a) calculada para todas las componentes principales puede escribirse como sigue:



[2]

lo que matricialmente, puede expresarse como $Z=X \cdot U$, siendo *Z* la matriz de valores de las componentes principales para cada caso y *U* la matriz cuyas columnas son respectivamente las direcciones de las componentes principales, ordenadamente.

Es interesante notar que la primera componente extraída explica más varianza que la segunda, ésta más que la tercera y así hasta la última, que será la que menos diversidad explique:

$$Var(Z_1) \geq Var(Z_2) \geq .... \geq Var(Z_p)$$

Obsérvese que incluso si la varianza de alguna de las componentes principales fuera nula, a partir de ella, todas las demás tendrían la misma propiedad; por lo que se podrían obviar ya que no explicarían ninguna variabilidad de los datos.

Pero ¿garantiza este procedimiento que no haya pérdida de información? Si la variabilidad total de los datos originales puede ser representada por la medida de varianza global $\sum_{i=1}^{p} \text{var}(X_i)$, parece lógico exigir que en el nuevo sistema coordenado ocurra que entre todas las componentes principales deben explicar toda esa misma variabilidad, y por tanto, la suma de todas las varianzas de las nuevas componentes debe ser igual a la varianza global que presentan los datos

$$\sum_{i=1}^{p} \text{var}(X_i) = \sum_{i=1}^{p} \text{var}(Z_i)$$

En realidad no será necesario imponer esta condición ya que, como veremos más adelante, tal y como se ha planteado el problema, esta propiedad, muy deseable, aparece como una consecuencia.

Obtención de la primera componente principal:

Dado que las nuevas componentes principales están centradas, por estarlo las originales, la varianza explicada por cualquier componente principal Z_h, h=l,...,p, puede expresarse como:

$$\text{var}(Z_h) = \frac{1}{n} Z_h' Z_h = \frac{1}{n} u_h' X' X u_h = u_h' (\frac{1}{n} X' X) u_h = u_h' S u_h \quad , \quad \forall_{h\,=\,1,\ldots,\,p}$$

siendo S la matriz de varianzas y covarianzas de los datos originales, por lo que el problema de obtener la primera componente principal puede expresarse como sigue:

$$\text{* Encontrar } u_1 \text{ tal que: } \quad |u_1| = 1 \quad y \quad \text{var}(Z_1) = \max_u \{\text{var}(Z)\}$$

o lo que es lo mismo,

$$\underset{u_1}{Max}\; u_1' S u_1 \quad \text{sujeto a}: u_1' u_1 = 1$$

Como es un problema de optimización con restricciones de igualdad, aplicamos el método de los multiplicadores de Lagrange; para lo que construimos su Lagrangiana L, obtenemos sus puntos estacionarios derivando respecto de las incógnitas e igualando a cero, y seleccionamos los máximos (Hessiano definido negativo).

$$L = u_1' S u_1 - \lambda \cdot (u_1' u_1 - 1)$$

$$\frac{\partial L}{\partial u_1} = 2 S u_1 - 2\lambda \cdot u_1 = 0 \Leftrightarrow S u_1 = \lambda u_1$$

$$\frac{\partial L}{\partial \lambda} = u_1' u_1 - 1 = 0 \Rightarrow \lambda = \lambda u_1' u_1 = u_1' \lambda u_1 = u_1' S u_1 = \text{var}(Z_1)$$

Por tanto, los posibles puntos (vectores, direcciones) estacionarios de la función que queremos maximizar —que se demuestra a través del Hessiano que son máximos— son

puntos que satisfacen las condiciones anteriores.

Recordemos que λ se dice autovalor (o valor propio) de una matriz A asociado a un autovector (o vector propio) u si y sólo si $Au=\lambda u$. Además, si A es una matriz cuadrada de dimensión p, simétrica y semidefinida positiva, tiene p-autovalores no negativos, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, a partir de los cuales se pueden extraer p-autovectores ortogonales.

Por tanto, las posibles soluciones de la igualdad $Su_1 = \lambda u_1$, son precisamente las parejas (λ,u) formadas por cada autovector u de esta matriz S y su correspondiente autovalor asociado λ. Así pues, las p posibles direcciones que en principio son candidatas a ser la primera componente principal son las marcadas por los p autovectores de la matriz S de varianzas y covarianzas.

De estas posibles soluciones, la primera componente principal será justamente aquélla que haga máxima la función $u_1' S u_1 = u_1' \lambda u_1 = \lambda u_1' u_1 = \lambda$, ya que $u_1' u_1 = 1$.

Así pues, la dirección de la primera componente principal será la del autovector asociado al mayor de todos los autovalores de la matriz S. Y la varianza explicada por esta primera componente será justamente su autovalor asociado.

Obtención del resto de componentes principales:

Procederemos por inducción y, una vez encontradas las h-1 primeras componentes principales, $1<h\leq p$, el problema será encontrar otra dirección normalizada u_h diferente y ortogonal a las direcciones de las anteriores componentes ya calculadas, y cuya varianza sea la máxima posible.

* Encontrar u_h tal que :

$|u_h| = 1$, ortogonal a los anteriores y con $\mathrm{V}ar(Z_h) = \max_{u_h}\{\mathrm{var}(Z_h)\}$

o lo que es lo mismo,

$$\underset{u_h}{Max}\ u_h' S u_h \quad \text{sujeto a : } u_h' u_h = 1 \quad \text{y} \quad u_h' u_j = 0\ ,\ \forall_{j<h}$$

Así que ahora la función Lagrangiana que debemos utilizar, puesto que tenemos otra maximización con restricciones de igualdad, sería la función a maximizar, menos λ por la primera restricción (la de normalidad), menos μ_1 por la segunda restricción (la de ortogonalidad con la primera componente principal), menos μ_2 por la tercera restricción (la de ortogonalidad con la segunda componente principal) y así sucesivamente; o sea:

$$L = u_h' S u_h - \lambda(u_h' u_h - 1) - \sum_{j=1}^{h-1} \mu_j u_h' u_j$$

Derivando con respecto de las incógnitas, obtenemos:

$$\frac{\partial L}{\partial u_h} = 2Su_h - 2\lambda \cdot u_h - \sum_{j=1}^{h-1} \mu_j \cdot u_j = 0$$
$$\frac{\partial L}{\partial \lambda} = u_h' u_h - 1 = 0$$
$$\frac{\partial L}{\partial \mu_j} = u_h' u_j = 0, \forall_{j<h}$$

Observemos que premultiplicando la primera ecuación por u_k', con k=1,2,...,h-1, y teniendo en cuenta las restricciones de normalidad y ortogonalidad de las componentes, y que para las anteriores componentes principales se cumple que $\lambda_k u_k = Su_k$, o equivalentemente dada la simetría de S que $\lambda_k u_k' = u_k' S$, obtenemos que:

$$2u_k' Su_h - 2\lambda \cdot u_k' u_h - \sum_{j=1}^{h-1} \mu_j \cdot u_k' u_j = 2\lambda_k u_k' u_h - 2\lambda \cdot u_k' u_h - \sum_{\substack{j=1 \\ j\neq k}}^{h-1} \mu_j \cdot u_k' u_j - \mu_k \cdot u_k' u_k = 0 \Leftrightarrow \mu_k = 0$$

En definitiva, $\mu_1 = \mu_2 = \cdots = \mu_{h-1} = 0$, y podríamos suprimir el sumatorio de la primera ecuación, quedando, de forma análoga a como obteníamos para la primera componente principal que:

$$2Su_h - 2\lambda u_h = 0$$

o equivalentemente,

$$Su_h = \lambda u_h$$

y de nuevo que:

$$\lambda = \lambda u_h' u_h = u_h' Su_h = \text{var}(Z_h)$$

En definitiva, las nuevas direcciones de la segunda, la tercera y las sucesivas componentes, también son direcciones marcadas por los autovectores de la matriz *S*, siendo las varianzas que explican iguales a sus autovalores asociados. Así, si $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ son los autovalores de *S* y la dirección de la primera componente principal era la del autovector asociado al máximo de los autovalores de la matriz S, λ_1, entonces la dirección de la segunda componente será la del autovector asociado al segundo autovalor en tamaño, λ_2; la tercera será la del autovector asociado al tercer autovalor en tamaño, y así sucesivamente, hasta la de la última componente que será la del autovector asociado al menor de los autovalores, λ_p; tomándose todos estos autovectores ortonormalizados, lo cual es siempre posible en una matriz semidefinida positiva como es la de varianzas y covarianzas S..

Si la matriz *S* tiene determinante cero, entonces tendrá al menos un autovector nulo y su correspondiente componente principal no recogerá nada de dispersión en su dirección.

Propiedades del ACP:

En lo sucesivo notaremos por Λ a la matriz diagonal de los autovalores de S, y por U a la matriz de sus autovectores asociados ortonormalizados y ordenadamente dispuestos en

columnas.

Nótese que, al estar los autovectores ortonormalizados, la matriz U es ortogonal. Es decir, verifica que su traspuesta coincide con su inversa, U'=U^{-1}, ya que $U'U = I$.

Además, la matriz diagonal Λ de los autovalores es semejante a la matríz S a través de la matriz *U*, por lo que sus trazas coinciden, ya que

$$Su_h = \lambda_h u_h, \forall h \Leftrightarrow SU = U\Lambda \Leftrightarrow U^{-1}SU = \Lambda \Leftrightarrow U'SU = \Lambda$$

Como de costumbre, notando por *X* la matriz de datos procedente de la observación de las variables originales (X_1, ..., X_p) y análogamente por *Z* la correspondiente matriz de valores calculados para las componentes principales (Z_1, ..., Z_p) para los mismos casos, podemos entonces expresar ésta, en función de *U* y *X,* como *Z=XU.*

Tras este preámbulo, destaquemos las siguientes propiedades de las Componentes Principales:

a) La matriz de varianzas y covarianzas de las componentes principales (Z_1, ..., Z_p) es:

$$Cov(Z) = U'SU = \Lambda \Rightarrow \text{var}(Z_h) = u_h^{'} S u_h = \lambda_h \quad \text{y} \quad \text{cov}(Z_h, Z_k) = 0 \quad (h \neq k)$$

La matriz de varianzas y covarianzas de las componentes principales (Z_1, ..., Z_p), puesto que éstas están centradas, puede calcularse como:

$$Cov(Z) = \frac{1}{n} Z'Z = \frac{1}{n}(XU)'(XU) = U'(\frac{1}{n} X'X)U = U'SU = \Lambda$$

En consecuencia, puesto que Λ es una matriz diagonal, las covarianzas de las componentes Z_l y Z_h son siempre nulas salvo cuando coinciden los índices, es decir, para la propia varianza de la componente correspondiente:

$$Cov(Z_l, Z_h) = \begin{cases} Var(Z_h) = u_h^{'} S u_h = \lambda_h & \text{si } l = h \\ 0 & \text{si } l \neq h \end{cases}$$

Por tanto, la varianza de la componente principal Z_h, coincide con el valor de su autovalor asociado λ_h.

b) La varianza total explicada por las componentes principales es la misma que la que presentan las variables originales. Es decir:

$$\sum_{i=1}^{p} Var(X_i) = Traza(S) = Traza(\Lambda) = \sum_{i=1}^{p} \lambda_h = \sum_{i=1}^{p} Var(Z_h)$$

Como en la diagonal principal de la matriz S aparecen las varianzas de las variables:

$$S = \begin{pmatrix} S_1^2 & Cov(X_1, X_2) & \cdots & \cdots \\ Cov(X_1, X_2) & S_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \cdots & S_p^2 \end{pmatrix}$$

entonces,

$$\sum_{i=1}^{p} Var(X_i) = Traza(S)$$

y como la matriz diagonal Λ de los autovalores es semejante a la matríz S a través de la matriz *U*, entonces sus trazas coinciden:

$$\sum_{i=1}^{p} Var(X_i) = Traza(S) = Traza(\Lambda) = \sum_{i=1}^{p} \lambda_h = \sum_{i=1}^{p} Var(Z_h)$$

c) Prop. de varianza explicada por $z_h = \dfrac{\lambda_h}{\sum_{i=1}^{p} \lambda_i}$

La proporción de varianza explicada por la componente Z_h se puede calcular como:

$$\frac{Var(Z_h)}{\sum_{i=1}^{p} Var(Z_h)} = \frac{\lambda_h}{\sum_{i=1}^{p} \lambda_i}$$

d) La matriz de covarianzas de las variables originales (X_1, ..., X_p) y las componentes principales (Z_1, ..., Z_p) es $Cov(X, Z) = SU = U\Lambda \Rightarrow \text{cov}(X_j, Z_h) = \lambda_h u_{jh}$

Matricialmente, basta con tener en cuenta que la matriz de covarianzas de las variables originales (X_1, ..., X_p) y las componentes principales (Z_1, ..., Z_p) es:

$$Cov(X, Z) = \frac{1}{n} X'Z = \frac{1}{n} X'(XU) = (\frac{1}{n} X'X)U = SU = U\Lambda$$

Pero veámoslo más explícitamente. Para calcular la covarianza de la variable X_j y la componente Z_h , como $U^{-1}=U'$ y $Z=XU$, multiplicando por la derecha por U^{-1}, tenemos equivalentemente que

$$ZU^{-1} = XUU^{-1} \quad \Leftrightarrow \quad ZU' = X$$

es decir, a través de esta expresión podemos poner las variables X_1,..., X_p, en función de las Componentes Principales y los autovectores. Así, sustituyendo X_j en $Cov(X_j , Z_h)$ por su correspondiente expresión según la fórmula anterior podremos calcular su valor:

$$X_j = \begin{pmatrix} Z_1 & \cdots & Z_p \end{pmatrix} \begin{pmatrix} u_{j1} \\ \vdots \\ u_{jp} \end{pmatrix}$$

pues la columna j-ésima de U' es la fila j-ésima de U, y por tanto,

$$X_j = \sum_{l=1}^{p} Z_l u_{jl}$$

y así, $Cov(X_j, Z_h) = Cov\left(\sum_{l=1}^{p} Z_l u_{jl} \ , \ Z_h \right) = \sum_{l=1}^{p} u_{jl} \ Cov(Z_l \ , Z_h)$

Y como las covarianzas de las componentes Z_l y Z_h son siempre cero, salvo cuando coincidan los índices, en cuyo caso sería la propia varianza de la correspondiente componente, esta expresión será distinto de cero cuando $l = h$, y en este caso, la covarianza entres la variable original X_j y la componente Z_h , será

$$Cov(X_j, Z_h) = Cov\left(\sum_{l=1}^{p} Z_l u_{jl} \ , \ Z_h \right) = \sum_{l=1}^{p} u_{jl} \ Cov(Z_l \ , Z_h) = u_{jh} Var(Z_h) = u_{jh} \cdot \lambda_h$$

ya que la varianza de Z_h es su autovalor asociado, λ_h.

e) El coeficiente de correlación lineal entre la variable X_j y Z_h (que denotaremos como r_{jh}) toma el valor: $r_{jh} = \frac{\sqrt{\lambda_h}}{\sigma_j} u_{jh}$

Sabemos que el coeficiente de correlación lineal es la covarianza dividida por el producto de sus respectivas desviaciones típicas:

$$r_{jh} = \frac{Cov(X_j \ , Z_h)}{S_j \ S_h}$$

Y sustituyendo los resultados anteriores, sería:

$$r_{jh} = \frac{Cov(X_j \ , Z_h)}{S_j \ S_h} = \frac{\lambda_h u_{jh}}{\sigma_j \sqrt{\lambda_h}} = \frac{\sqrt{\lambda_h}}{\sigma_j} u_{jh}$$

f) Estando $X_1, ..., X_p$ tipificadas $\Rightarrow r_{jh} = \sqrt{\lambda_h} u_{jh}$ y la varianza explicada por $Z_h = \frac{\lambda_h}{p}$

Si las variables $X_1, \cdots, X_p$ están tipificadas, la desviación típica de cada variable original sería igual a la unidad, de donde sustituyendo en la expresión de la propiedad anterior obtenemos la expresión particular del coeficiente de correlación.

Obsérvese que si las variables iniciales están tipificadas, los coeficientes que conforman las direcciones de las componentes principales están muy relacionados con los coeficientes de correlación lineal de éstas con las variables originales.

Dada la invarianza del coeficiente de correlación ante cambios de origen y escala, este mismo valor de correlación se obtienen cuando tipificamos las componentes principales, Z_h (que son centradas), dividiéndola por su desviación típica , $\sqrt{\lambda_h}$ (ya que su varianza es el autovalor), obteniendo la componente tipificada, Y_h:

$$Z_h = u_{1h}X_1 + u_{2h}X_2 + \cdots + u_{ph}X_p \Rightarrow Y_h = \frac{u_{1h}}{\sqrt{\lambda_h}}X_1 + \frac{u_{2h}}{\sqrt{\lambda_h}}X_2 + \cdots + \frac{u_{ph}}{\sqrt{\lambda_h}}X_p$$

cumpliéndose, por tanto, que

$$\text{Si } X_j \text{ está tipificada} \quad \Rightarrow \quad corr(X_j, Y_h) = corr(X_j, Z_h) = \sqrt{\lambda_h}u_{jh}$$

Por otra parte, y también en el caso de variables tipificadas, obtenemos que

$$\sum_{i=1}^{p}\lambda_j = \sum_{i=1}^{p}Var(Z_j) = \sum_{i=1}^{p}Var(X_j) = 1 + \cdots + 1 = p$$

de donde sustituyendo en la propiedad c), la proporción de varianza explicada por Z_h es justamente el cociente λ_h partido por *p*:

$$\frac{\lambda_h}{\sum_{i=1}^{p}\lambda_i} = \frac{\lambda_h}{p}$$

Selección del número de factores:

En la introducción al tema dijimos que esta técnica del Análisis de Componentes Principales iba a permitir, generalmente, la reducción de la complejidad del problema original mediante la reducción de la dimensión del mismo.

Ello será inmediato cuando exista alguna dependencia perfecta entre algunas de las variables originales, en cuyo caso la matriz de varianzas y covarianzas presentará al menos un autovalor nulo. En este caso, existirán tantas componentes principales despreciables (desde el punto de vista de que no explican nada de la diversidad observada) como autovalores nulos haya. Pero cuando no haya ninguna dependencia perfecta entre algunas de las variables originales, todas las Componentes Principales aportan información de forma incorrelacionada con las otras. Por ello, si queremos representar con las componentes principales toda la diversidad manifiesta en los datos originales sin ninguna pérdida de información, estrictamente no deberíamos eliminar a ninguna.

Sin embargo, no todas las componentes principales aportan la misma información. Y como

las hemos extraído ordenadamente de más a menos explicativas, podemos esperar que las últimas aporten muy poco al explicar muy poca varianza de la diversidad de los datos que tenemos. Así que, a nivel práctico, podremos decidir cuándo quedarnos con un conjunto reducido de componentes (obviamente con un número reducido de las primeras componentes principales), despreciando las últimas a sabiendas de que las componentes retenidas no van a explicar totalmente la diversidad original (varianza total), pero sí gran parte de ella.

El problema que se plantea es pues cómo decidir con cuantas componentes nos quedamos. Veamos los tres principales criterios utilizados para ello que, en cualquier caso, no deben ser nunca entendidos como alternativos sino como complementarios.

Criterio de la media:

Este **criterio de la media** se basa en el siguiente razonamiento intuitivo: cada una de las variables originales nos proporcionan información sobre la diversidad de comportamientos de acuerdo con la varianza que presenta (a mayor varianza, mayor diversidad de comportamientos y más informativa es por tanto la variable; y viceversa). Por término medio, un indicador de la información que aporta cada variable original, en términos de dispersión, es la varianza media. Parece lógico pensar, pues, que cuando una componente principal explique más de lo que explica en promedio una de las variables originales, aquélla está contribuyendo más al conocimiento del comportamiento de los datos que lo que lo hacen algunas de las variables originales. Por ello, este criterio nos induce a actuar reteniendo las componentes principales que expliquen más que la media de las variables originales y a desechar aquéllas componentes que expliquen menos que la media, ya que estarían explicando en términos medios menos de lo que lo hacía una variable original. Dado que la varianza explicada por una componente principal es su autovalor asociado λ_h, el criterio de la media consiste, por tanto, en compararlo con la varianza media de las variables originales, que sabemos coincide con la media de todos los autovalores de la matriz S, $\overline{\lambda}$.

Así pues, con el criterio de la media retendremos para nuestros estudios las *h* primeras componentes, donde *h* es tal que verifica:

$$\lambda_h > \overline{\lambda} = \frac{1}{p}\sum_{i=1}^{p}\lambda_i \qquad \text{y} \qquad \lambda_{h+1} < \overline{\lambda}$$

Si las variables están tipificadas, las variables originales tendrían varianza 1 y su promedio sería igualmente 1. En este caso, el criterio de la media compara los autovalores de la matriz S con la unidad y decide la retención o no de las componentes con la misma lógica expuesta anteriormente. Es decir, retendremos para nuestros estudios las *h* primeras componentes, donde *h* es tal que:

$$\lambda_h > 1 \quad \text{pero} \quad \lambda_{h+1} < 1$$

En la práctica, tomar exactamente la media de las variables originales (1 si están tipificadas) como umbral de corte para decidir si la retenemos o no, presenta un problema de asimetría que hace esta opción demasiado exigente, en general. El problema de asimetría al que nos referimos viene motivado por el hecho de que las componentes principales están

incorrelacionadas (aportan información neta sobre la diversidad), mientras que las variables originales suelen presentar correlaciones entre ellas (pueden compartir o repetir información con otras variables). Por ello comparar si la información (neta) aportada por una componente principal es estrictamente superior a la información promedio proporcionada por las variables originales (probalemente inflada por las duplicidades producida por las correlaciones) parece una decisión sesgada en contra de la mayor calidad informativa de las componentes principales; por lo que el valor umbral de comparación suele relajarse, en función de las correlaciones observadas, siendo un valor estandar generalmente aceptado el 0,8. Así, el criterio quedaría como retener para nuestros estudios las *h* primeras componentes, sempre que:

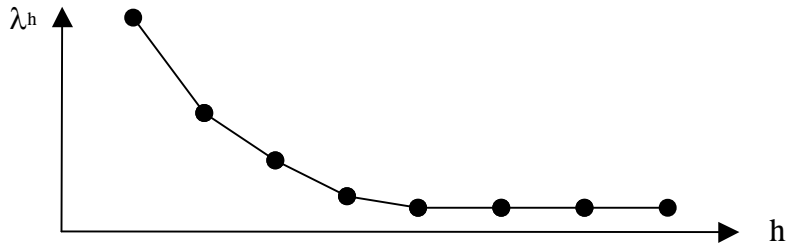
$$\lambda_h > 0.8\overline{\lambda} \qquad \text{y} \qquad \lambda_{h+1} < 0.8\overline{\lambda}$$

y en el caso de que las variables originales estuviesen tipificadas, siempre que:

$$\lambda_h > 0.8 \quad \text{pero} \quad \lambda_{h+1} < 0.8$$

Criterio del gráfico de sedimentación:

Este criterio utiliza la representación gráfica del decaimiento de la varianza explicada por las sucesivas componentes principales. Así, representando ordenadamente las componentes principales en el eje de abscisas mediante números enteros (1, 2, ····, *p*) que las representan, y haciéndoles corresponder a cada uno su correspondiente autovalor (λ_1 ,λ_2 ,····,λ_p) en las ordenadas, al estar éstos necesariamente ordenados por el propio método de extracción del ACP ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$), obtendremos un gráfico generalmente decreciente (realmente es monótono no creciente ya que algunos autovalores podrían repetirse) y en el que puede generalmente apreciarse que los decrecimientos rápidos se produce entre los primeros autovalores y que, a partir de cierto orden de la componente principal, el gráfico empieza a estabilizarse tomando un aspecto similar al siguiente:



Este perfil típico se asemeja al de un valle formado por el efecto de la erosión y sedimentación geológicas, de donde recibe el nombre; aunque también se suele conocer a este gráfico como “gráfico de codo” por su parecido intuitivo con el mismo.

Puesto que la zona en la que se estabiliza el gráfico suele ser la de valores pequeños de autovalores asociados a las últimas componentes principales, y por tanto menos importantes para la representación de la diversidad, el criterio del gráfico de sedimentación consiste en quedarse justamente con las componentes principales previas a la zona de sedimentación. Esto nos llevaría, sobre el gráfico esquemático anterior, a quedarnos con las tres primeras componentes principales, pues a partir de la cuarta comienza la zona de sedimentación y los

correspondientes autovalores son relativamente pequeños.

Claro, que esta decisión siempre debe ser corroborada por el carácter realmente pequeño de las varianzas (autovalores) de las componentes desechadas.

Criterio de la proporción conveniente de varianza explicada

Aunque la reducción es un objetivo deseable, no lo es menos el realismo o verosimilitud de la información finalmente retenida mediante cualquier mecanismo o criterio de reducción de la información. Generalmente, simplificación de un problema o la reducción de su dimensionalidad conllevan la perdida de cierta cantidad de información. Por ello siempre debemos tener en cuenta que esta pérdida de información sea tolerable.

El criterio de la proporción conveniente de varianza explicada trata de garantizar este extremo, para lo que se fijará un umbral de información, U, en términos de proporción de la varianza total explicada, por debajo del cual no está dispuesto a bajar. Como la suma ordenada de los primeros autovalores coincide con la varianza explicada por las respectivas primeras componentes principales, este criterio nos hará retener tantas componentes, de las primeras, como fueran necesarias para alcanzar conjuntamente el umbral deseado.

Por tanto, el criterio de la proporción conveniente de varianza explicada, fijado un umbral para la proporción de varianza explicada, U, nos dice que retengamos para nuestros estudios las *h* primeras componentes, donde *h* es tal que verifica:

$$\frac{\sum_{i=1}^{h}\lambda_i}{\sum_{i=1}^{p}\lambda_i} > U \quad \text{y} \quad \frac{\sum_{i=1}^{h-1}\lambda_i}{\sum_{i=1}^{p}\lambda_i} < U$$

Interpretación de las componentes principales

La problemática de la interpretación de las componentes principales es, nada más y nada menos, tratar de asignar un significado inteligible y útil a las componentes principales obtenidas.

Para ello se recurre a examinar la relación existente entre las componentes principales y las variables originales (u otras auxiliares), para por medio de esta relación tratar de darles un contenido a su significado, para lo que la información básica para esta tarea es la matriz de correlaciones entre las componentes principales y las variables originales, que toma en este ámbito el nombre de *matriz de estructura*.

Representaciones gráficas

Como mecanismo auxiliar para facilitar este análisis interpretativo se suelen construir

gráficos para representar las variables originales en espacios de referencia de las componentes principales, que muestran las proyecciones de la nube de puntos p-dimensional sobre los subespacios formados por, generalmente, las 2 ó 3 componentes principales y en los que éstas serían los nuevos ejes de referencia.

Análogamente, para permitir caracterizar mejor a los casos de acuerdo con las nuevas componentes principales, también suelen representarse a los casos en el espacio de referencia de las componentes principales.