DOI: 10.22199/S07160917.1984.0008.00005

REVISTA PROYECCIONES N° 8: 109-136
Diciembre 1984 - ISSN 0716-0917
JORNADA MATEMATICAS, Agosto 1984.

USO DE COMPONENTES PRINCIPALES EN DATOS COMPOSICIONALES

JORGE GALBIATI RIESCO*

---

1. EL PROBLEMA DE LAS COMPONENTES PRINCIPALES.

Cuando se dispone de datos multidimensionales, es deseable reducir su dimensionalidad, a cambio de perder una parte de la información. Esto es debido a que es muy difícil representar y entender la información contenida en datos de dimensiones mayores que 2 ó 3. La reducción de dimensionalidad puede hacerse de diferentes formas, apuntando todas ellas a que la pérdida de información, por la simplificación que se hace, sea mínima.

En el análisis en componentes principales, se centra la atención en la varianza, pues una mayor varianza permite separar más las observaciones y así poder discriminar mejor a los individuos que confor-

---

* Profesor Departamento Matemáticas, Universidad Católica de Valparaíso.

man la muestra. Desde este punto de vista, la varianza viene a ser una medida de la cantidad de información disponible. Cuando en alguna dimensión hay una varianza pequeña, los datos se agrupan en lo que puede considerarse un sólo conglomerado, en esa dimensión, la que tiende a ser desechada del análisis, por aportar poca información.

Supongamos que se tiene un vector aleatorio X, p-dimensional, esperanza $E(X)=\mu$, matriz de varianzas-covarianzas $Var(X) = \sum = (\delta_{ij})$. El vector X representa la información p-dimensional. Convencionalmente losvectores V como columnas, en términos de sus coordenadas.

Interesa encontrar una dirección en el espacio R^p, tal que la proyección de X sobre ella tenga varianza máxima.

Es decir, se debe encontrar un vector $A =(A_1,A_2,....,A_p)$, tal que

$$Y= A'X \qquad (1)$$

tenga varianza máxima, sujeto a la condición de normalización

$$A'A = 1 \qquad (2)$$

para independizar el resultado de la magnitud de A.

Se tiene que

$$Var(Y)= Var(A'X) = A'Var\ (X)\ A = A'\Sigma A \qquad (3)$$

Entonces se debe maximizar $A'\Sigma A$, sujeto a la condición $A'A= 1$.

Derivando la función

$$\Phi(A,\lambda) = A'\Sigma A + \lambda(1 - A'A)$$

$$\Phi(A,\lambda) = \sum_{k,1} A_k \delta_{k1} A_1 + \lambda (1 - \sum_k A_k^2) \qquad (4)$$

respecto de los componentes de A e igualando a cero las derivadas para encontrar puntos críticos, se obtiene lo siguiente, considerando la simetría de Σ :

$$\frac{\partial \Phi}{\partial A_j} = \sum_1 \delta_{j1} A_1 + \sum_k A_k \delta_{kj} - 2\lambda A_j = 0 \qquad (j = 1,2,.....p)$$

$$\frac{\partial \Phi}{\partial A_j} = 2 \sum_1 \delta_{j1} A_1 - 2\lambda A_j = 0$$

En forma vectorial, se puede expresar

$$\frac{\partial O}{\partial A} = 2 \Sigma A - 2 \lambda A = 0 \qquad (5)$$

Derivando respecto del multiplicador de Lagrange λ ,

$$\frac{\partial \Phi}{\partial \lambda} = 1 - \sum_k A_k^2 = 0 \qquad (6)$$

De la ecuación (5),

$$\Sigma A = \lambda A \qquad (7)$$

La varianza, entonces, es

$$Var(Y) = A'\Sigma A = \lambda A'A = \lambda \qquad (8)$$

2. ALGUNAS PROPIEDADES MATRICIALES.

DEFINICION.

Dada una matriz Σ , un vector no nulo v tal que cumple la ecuación

$$\Sigma v = \lambda v$$

en que λ es un escalar, se denomina vector propio de Σ. El escalar λ se llama valor propio de Σ.

Definición: Sea Σ una matriz simétrica real de orden p x p

a) Σ es semidefinida positiva si dado $v \in R^p$, $v \neq 0$, se cumple que $v' \Sigma v \geq 0$.

b) Σ es definida positiva si dado $v \in R^p$, $v \neq 0$, se cumple que $v' \Sigma v > 0$.

Proposición: Sea Σ una matriz simétrica real. Entonces sus valores propios son reales.

Proposición: Sea Σ una matriz simétrica, real, de orden p x p. Σ tiene p vectores propios, que forman una base ortogonal de $\mathbb{R}^p$, si Σ no sing.

Proposición: Los valores propios de una matriz definida positiva son positivos.

Proposición: Si Σ es una matriz simétrica real de orden p x p, existe una matriz ortogonal α tal que

$$\alpha' \Sigma \alpha = D$$

con $D = \text{diag}\{\lambda_1, \lambda_2, \ldots\ldots, \lambda_p\}$

Los λ_i son los valores propios de Σ y $\alpha = [a_1, a_2, \ldots, a_p]$, matriz cuyas columnas son los vectores propios de Σ, ordenados de acuerdo a los valores propios en D.

Definición: Si dos matrices A y B están relacionadas por una ecuación del tipo

$$U^{-1} AU = B$$

para alguna matriz U no singular, se dice que son similares.

En nuestro caso Σ es matriz similar a una matriz diagonal D. Se dice entonces, que Σ esdiagonizable. Además se da la particularidad de que la matriz simétrica es diagonizable a través de una matriz ortogonal, pues α es ortogonal ($\alpha'\alpha = \alpha\,\alpha' = I$)

Proposición: Si Σ y D se relacionan de acuerdo a la ecuación $\alpha'\Sigma\alpha = D$, entonces

$$\text{traza}\,(\Sigma) = \text{traza}\,(D) = \sum_{1=1}^{p} \lambda_i$$

y

$$\det(\Sigma) = \det(D) = \prod_{1=1}^{p} \lambda_i$$

3. SOLUCION AL PROBLEMA DE LAS COMPONENTES PRINCIPALES.

Supondremos que la matriz de varianza-covarianza Σ es definida positiva, por construcción.

De acuerdo con la definición, y por la ecuación (7), se ve que la solución al problema de las componentes principales está dada por los vectores propios normalizados (de norma 1) de la matriz de varianza-covarianza Σ .

Como las varianzas obtenidas son los correspondientes valores propios (8), queda como solución aquel vector propio correspondiente al mayor de los valores propios, y que simbolizaremos A_1 y λ_1 respectivamente.

Corresponde a un máximo absoluto, pues si V es otro vector no nulo de $\mathbb{R}^p$, de norma 1, y como los vectores propios $A_1, A_2, \ldots, A_p$ forman una base ortogonal, V puede expresarse como combinación lineal de ellos:

$$V = \sum_{1=1}^{p} c_i \, A_i \qquad (9)$$

V tiene norma 1, luego

$$1 = V'V = \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j A_i A_j = \sum_{i=1}^{p} C_i^2 \qquad (10)$$

por la ortogonalidad de los A_i.

Entonces $Var(V'X) = V'Var(X)V = V'\Sigma V = \sum_{i=1}^{p} c_i A_i' \; \Sigma \sum_{j=1}^{p} c_j A_j =$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j A_i' \; \Sigma \, A_j = \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j A_i' \; \Sigma \, A_j =$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j \, \lambda_i \, A_i' \, A_j = \sum_{i=1}^{p} c_i^2 \, \lambda_i \qquad (11)$$

por la ortogonalidad de los A_i.

De (10) y (11) resulta que Var(V'X) es un promedio ponderado de los λ_i, cuyas ponderaciones C_i^2 suman 1, por lo que es menor o igual que el mayor de los λ_i, λ_1.

Por lo tanto este último es un máximo absoluto de la función $Y = A'X$ sujeto a la restricción $A'A = 1$, y se alcanza este máximo cuando A es igual al vector propio A_1 asociado a él.

4. EL SISTEMA DE EJES PRINCIPALES.

DEFINICION.

La proyección del vector aleatorio X sobre el vector A_1, solución del problema de las componentes principales, se denomina Primera Componente Principal. La dirección en el espacio $\mathbb{R}^p$ definida por A_1, se denomina Primer Eje Principal.

Hemos visto que la varianza de X proyectado sobre A_1, $Y = A'_1X$, es $A'_1 \Sigma A_1$, en que Σ es la matriz de varianza-covarianza de X. Y es la máxima sobre las varianzas de todas las proyecciones posibles de X sobre vectores de norma 1.

A continuación supondremos los vectores propios A_1, $A_2,.....A_p$ ordenados según el orden de sus respectivos valores propios λ_1, $\lambda_2,....., \lambda_p$, de mayor a menor.

Definición: Dado un vector aleatorio X, su k-ésima componente principal es su proyección sobre el vector propio normalizado correspondiente al valor propio de orden k, ordenados de mayor a menor ($1 \leq k \leq p$). El k-ésimo eje principal es la dirección definida por este vector.

Proposición: Sean $A_1, A_2, ..., A_p$ los vectores propios de X, ordenados de acuerdo a sus respectivos valores propios en orden de mayor a menor. Sea k un entero tal que $1 \leq k \leq p$, y sea S_k el subespacio de $\mathbb{R}^p$ generado por los vectores propios A_k, $A_{k+1}, ..., A_p$.

Las proyecciones de X sobre un vector normalizado de S_k tiene máxima varianza cuando ese vector es A_k.

La demostración es similar a la demostración de que la varianza máxima se obtiene con A_1, haciendo las sumas desde k hasta p.

5. ESTRUCTURA DE VARIANZAS.

Las componentes principales, entonces, definen direcciones ortogonales en las que al proyectar el vector aleatorio X, se obtiene varianzas máximas, cada una ortogonal a las restantes. Con esto se forma un nuevo sistema de coorde

nadas cuyas direcciones, en el espacio original, están dadas por $A_1, A_2, \ldots\ldots, A_p$, en ese orden. Los valores de las varianzas son respectivamente $\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$, en orden decreciente de magnitud.

La matriz que transforma las coordenadas de los vectores del sistema de ejes a los ejes principales es $\mathbb{A}'$, en que

$$\mathbb{A}' = \begin{bmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{bmatrix} \qquad (12)$$

cuyas filas son los vectores propios.

En efecto, si V es un vector de $\mathbb{R}^p$, en las coordenadas principales V se expresa

$$\mathbb{A}'V = \begin{bmatrix} A_1'V \\ A_2'V \\ \vdots \\ A_p'V \end{bmatrix} \qquad (13)$$

siendo cada coordenada las respectiva componente principal.

Además $\mathbb{A}$ es precisamente la matriz ortogonal que diagonaliza Σ , pues por estar constituída por vectores propios,

$$\Sigma\, \mathbb{A} = D\, \mathbb{A}, \text{ con } D = \text{diag}\{\lambda_1, \lambda_2, \ldots\ldots, \lambda_p\}$$

y como es ortogonal,

$$\mathbb{A}'\Sigma\, \mathbb{A} = D \qquad (14)$$

La varianza total de $X = (x_1, x_2, \ldots, x_p)$, es

$$\sum_{i=1}^{p} Var(x_i) = \sum_{i=1}^{p} \delta_{ii} = \text{traza } (\Sigma) \qquad (15)$$

Si $Y = \mathbb{A}'X = \begin{bmatrix} A_1'X \\ A_2'X \\ \vdots \\ A_p'X \end{bmatrix}$ es el vector aleatorio X referido a

las coordenadas especiales, la varianza total de Y es

$$\sum_{i=1}^{p} Var(A_1'X) = \sum_{i=1}^{p} \lambda_i = tr(D) \qquad (16)$$

La proporción de varianza de la componente principal k (varianza explicada por esa componente) es

$$\frac{Var(A_k'X)}{\text{Varianza total}} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} = \frac{\lambda_k}{\text{traza } (D)} \qquad (17)$$

La proposición de varianza explicada por las primeras k componentes principales es

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad (18)$$

Como $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, dada una proposición de varianza $q<1$, se puede encontrar un número entero $k \leq p$ tal que

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_{k-1}}{tr\ (D)} < q \leq \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{tr\ (D)}$$

Si k<p, el subespacio generado por las primeras k componentes principales explica al menos la proporción q de la varianza total del vector aleatorio Y (o X).

En tal caso se ha logrado reducir la dimensión de la variable aleatoria que representa la información, con lo que se ha logrado simplificar el problema. El costo ha sido una pérdida de información, que en todo caso se ha minimizado, y que no es mayor que la fracción 1-q de la varianza total. Se ha minimizado precisamente porque las componentes principales han maximizado la varianza proyectada en los ejes.

6. ESTRUCTURA DE COVARIANZAS.

La matriz de varianzas-covarianzas de X es Σ, por lo que $cov(x_i,x_j) = \delta_{ij}$, el elemento (i,j) de la matriz.

La covarianza entre las proyecciones sobre dos ejes principales distintos es

$$cov(y_i,y_j) = cov(A_i'X, A_jX) = A_i' Cov(X,X') A_j = A_i' Var(X) A_j =$$

$$= A_i' \Sigma A_j = \lambda_j A_i' A_j = 0 \qquad (19)$$

si $i \neq j$, por ortogonalidad.

Este resultado, de que las covarianzas entre proyecciones de X sobre los ejes principales es cero, ya se podía haber deducido, pues

$$Var(Y) = Var(\mathbb{A}'X) = \mathbb{A}'Var(X)\,\mathbb{A} = \mathbb{A}'\Sigma\,\mathbb{A}$$
$$Var(Y) = D = diag\{\lambda_1, \lambda_2, \ldots, \lambda_p\} \qquad (20)$$

Por último, la covarianza entre X e Y es

$$Cov(X,Y) = Cov(X', \mathbb{A}'X) = Cov(X'X)\,\mathbb{A} = Var(X)\,\mathbb{A} = \Sigma\,\mathbb{A}$$

Pero $\mathbb{A}'\Sigma\,\mathbb{A} = D$ y $\mathbb{A}$ es ortogonal; luego

$$\Sigma\,\mathbb{A} = \mathbb{A}D$$
$$Cov(X,Y) = \mathbb{A}D$$

Con lo que

$$cov(x_i, y_j) = (\mathbb{A}D)_{ij} = A_{ij}\,\lambda_j \qquad (21)$$

La correlación es

$$corr(x_i, y_j) = \frac{A_{ij}\,\lambda_j}{\sqrt{var(x_i)\,var(y_j)}} = \frac{A_{ij}\,\lambda_j}{\sqrt{\delta_{ii}\,\lambda_j}} =$$

$$corr(x_i, y_j) = A_{ij}\sqrt{\frac{\lambda_j}{\delta_{ii}}} \qquad (22)$$

7. SITUACION MUESTRAL.

Ahora supongamos que disponemos de una muestra de n observaciones provenientes de un vector aleatorio X p-dimensional.

Usaremos el símbolo $\mathbb{X}$ para denotar la matriz de datos de orden nxp, cuyas filas $x_k = (X_{k1}, x_{k2}, \ldots . x_{kp})$ son las observaciones p-dimensionales y las columnas $x_{(j)}$ las variables o características medidas. Supondremos n > p.

La matriz de varianzas-covarianzas Σ se estima por la matriz de varianzas-covarianzas muestral,

$$S_{pxp} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})' = \frac{1}{n} \sum_k x_k x_k' - \bar{x}\bar{x}' = (S_{ij}) \quad (23)$$

$$\text{con } \bar{x}_{px1} = \frac{1}{n} \sum_{k=1}^{n} x_k \quad \text{vector de medias} \quad (24)$$

Las propiedades que se refieren a valores propios se mantienen, pues S es simétrica, semidefinida positiva, aunque se supondrá que es definida positiva, pues este es el caso normal. Para que no lo sea, tendría que tener rango menor que p. Esto significa que las columnas de X son linealmente dependientes, lo que es muy poco probable por ser una muestra aleatoria, por lo tanto es razonable suponer que no ocurra esta situación.

Las componentes principales, en forma natural, se estiman utilizando los vectores propios normalizados a_i de S, y se ordenan de acuerdo al orden descendente de magnitud de sus respectivos valores propios 1_i.

En tal caso se dan las siguientes propiedades:

Sea $\alpha = [a_1, a_2, \ldots, a_p]$ la matriz cuyas columnas son los vectores propios de S ordenados según los valores propios.

$$\text{Entonces } \alpha' S \alpha = L \quad (25)$$

con $L = \text{diag}\{1_1, 1_2, \ldots, 1_p\}$
y α ortogonal.

L es la matriz de varianzas-covarianzas de la matriz

de datos transformados al nuevo sistema de coordenadas principales.

$$\mathbf{Y} = \mathbf{X}\,\alpha$$

$$\mathbf{Y} = (y_{ij})$$

y_{ij} representa la magnitud de la proyección de la observación i-ésima x_i sobre el j-ésimo vector propio a_j.

Por lo tanto la varianza de la j-ésima variable de Y es l_j, y la covarianza entre dos variables distintas es cero.

La proporción de varianza contenida en la proyección de la variable j-ésima es

$$\frac{l_j}{l_1 + l_2 + \ldots + l_p} = \frac{l_j}{\text{traza } (L)} \qquad (26)$$

y la proporción de la varianza contenida en las proyecciones de las k primeras variables es

$$\frac{l_1 + l_2 + \ldots + l_k}{l_1 + l_2 + \ldots + l_p} \qquad (27)$$

La covarianza entre $\mathbf{X}$ e $\mathbf{Y}$ está dada por la matriz de covarianza

$$\text{Cov}(\mathbf{X},\mathbf{Y}) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})' = \frac{1}{n} \sum_i x_i y_i - \bar{x}\bar{y}'$$

en que x_i e y_i son las filas i-ésimas de $\mathbf{X}$ e $\mathbf{Y}$ respectivamente, y

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{e} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

vectores de medias.

Pero $\mathbb{Y} = \mathbb{X}\alpha$ luego

$$\mathrm{Cov}(X,Y) = \frac{1}{n} \sum_{k=1} (x_k - \bar{x})\,(x_k - \bar{x})'\alpha = S\alpha \qquad (28)$$

Además $\alpha' S \alpha = L$ y α es ortogonal, luego

$$S\,\alpha = \alpha\,'L$$

La covarianza entre una variable original $x_{(i)}$ y otra variable $y_{(j)}$ en un eje principal, es entonces

$$\mathrm{cov}(x_{(i)}, y_{(j)}) = (\alpha'L)_{ij} = a_{ij}\ l_j \qquad (29)$$

Su correlación es

$$\mathrm{corr}(x_{(i)}, y_{(j)}) = \frac{a_{ij}\ l_j}{\sqrt{\mathrm{var}(x_{(i)})\mathrm{var}(y_{(j)})}} = \qquad (30)$$

$$= \frac{a_{ij}\ l_j}{\sqrt{s_{ii}\ l_j}} = a_{ij}\sqrt{\frac{l_j}{s_{ii}}}$$

8. COMPONENTES PRINCIPALES GENERALIZADAS.

La transformación de componentes principales $\alpha : \mathbb{R}^p \to \mathbb{R}^p$ definida anteriormente, y que transforma una matriz de datos $\mathbb{X}$ en otra $\mathbb{Y} = \mathbb{X}\,\alpha$ con las características indicadas, es una transformación lineal de los datos.

Tiene por lo tanto, el inconveniente de preservar la forma que tiene el conjunto de datos en el espacio $\mathbb{R}^p$. Si el

conjunto presenta algún tipo de curvatura, una transformación lineal entregará coordenadas que no podrán adaptarse a ella, perdiéndose la posibilidad de reducir aún más la dimensionalidad de la información.

Esto sí se puede lograr utilizando coordenadas curvilíneas adecuadas. El procedimiento es el siguiente:

El vector aleatorio original es

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

y se dispone de un conjunto de k funciones conocidas (posiblemente no lineales) de X, $f_1(X)$, $f_2(X)$,...., $f_k(X)$, $k \geq p$, linealmente independientes. k no debe ser menor que p, con el objeto de preservar la cantidad de información.

Entonces se dispone de un nuevo vector aleatorio

$$F = \begin{bmatrix} f_1(X) \\ f_2(X) \\ \vdots \\ f_k(X) \end{bmatrix} \tag{31}$$

al cual se le aplicará un análisis en componentes principales como el indicado anteriormente. Es decir, se buscarán vectores ortogonales normalizados A_1, A_2,...., A_k, tales que las combinaciones lineales

$$A_i' F = \sum_{j=1}^{k} a_{ij} f_j \quad (x_1, x_2, \ldots, x_p) \tag{32}$$

$i = 1, 2, \ldots, k$

tengan máxima varianza el primero, y máxima varianza cada uno de los siguientes en el subespacio ortogonal a los anteriores vectores propios A_i.

Como en el caso lineal, estos vectores son los vectores propios de la matriz varianzas-covarianzas de F, Σ_F.

En el caso de una muestra, se trabaja en forma análoga, transformando primero los datos, y luego efectuando el análisis como se indica en el punto 7.

9. DATOS COMPOSICIONALES.

Los datos composicionales son vectores de proporciones que componen un todo. Por lo tanto tienen la restricción de que sus coordenadas suman 1.

Por ejemplo, en geología, las proporciones de diversos compuestos determinados, que forman un tipo de roca, constituyen un dato composicional. La composición, en proporciones de nacionalidades de un grupo humano. Las proporciones de respuestas por alternativas en una pregunta de alternativas múltiples, en una encuesta.

Veamos algunas definiciones y notaciones relacionadas con datos composicionales:

Designaremos $\mathbb{P}^p$ al octante positivo de $\mathbb{R}^p$,

$$\mathbb{P}^p = \{ (x_1, x_2, \ldots, x_p) \;/\; x_i > 0, \quad i = 1, 2, \ldots, p \} \tag{33}$$

Designaremos S^p al simplex positivo de $\mathbb{R}^p$,

$$S^p = \{(x_1, x_2, \ldots, x_p)' / x_1 + x_2 + \ldots + x_p < 1,\ x_i > 0,\ i = 1,2,\ldots p\} \tag{34}$$

Y designaremos $\mathbb{H}^{p+1}$ a la región del octante positivo de $\mathbb{R}^p$ del hiperplano cuya ecuación es $x_1 + x_2 +,\ldots., + x_p = 1$,

$$\mathbb{H}^{p+1} = \{(x_1, x_2,\ldots.,x_{p+1}/x_1 + x_2+\ldots.+ x_{p+1} = 1,\ x_i > 0,\ i=1,2,..,p+1\} \quad (35)$$

DEFINICION.

Un vector o punto de $\mathbb{H}^{p+1}$ se llamará composición, y un conjunto de tales vectores, datos composicionales.

Se usará la notación $X^{(p)}$ para enfatizar la dimensión de X.

Dado un vector $X^{(p)} \in S^p$, $X^{(p)} = (x_1, x_2,\ldots., x_p)'$, se puede construír una composición en forma natural, $X^{(p+1)} \in \mathbb{H}^{p+1}$ haciendo $x_{p+1} = 1- x_1 -x_2-\ldots. - x_p$, y $X^{(p+1)} = (x_1, x_2,\ldots., x_p, x_{p+1})'$, el vector $X^{(p)}$ aumentado agregando la coordenada x_{p+1}.

Si $c<p$, $X^{(c)} = (x_1, x_2,\ldots., x_c)'$ es un subvector de $X^{(p)} = (x_1, x_2,\ldots., x_p)'$.

$T(X^{(c)})$ denotará la suma $T(X^{(c)}) = x_1 + x_2 +\ldots. + x_c$

Consideremos la función

$C : \mathbb{P}^{p+1} \to S^p$ definida por

$X^{(p)} = C(W^{(p+1)})$ en que

$$x_i = \frac{w_i}{T(W^{(p+1)})} \qquad \frac{w_i}{w_1 + w_2 +\ldots.+ w_{p+1}} \qquad i= 1, 2,\ldots., p$$

Definición: El vector $W^{(p+1)}$ tal que $X^{(p)} = C(W^{(p+1)})$ se llama base de $X^{(p)}$.

De la definición anterior, se ve claramente que dado un vector $X^{(p)}$, su base no es única. En efecto, si W es una base de X, bW también lo es, en que b es un escalar positivo cualquiera.

10. DIFICULTADES EN EL TRATAMIENTO DE DATOS COMPOSICIONALES.

Los datos composicionales presentan dos características que es necesario hacer notar.

La primera es que la suma de sus coordenadas es 1. Esta restricción implica que pertenecen a un subespacio de dimensión menor en una unidad añ número de coordenadas. Por lo tanto la matriz de varianzas-covarianzas tiene rango una unidad menor que el número de filas.

El análisis en componentes principales busca producir ejes ortogonales sobre los cuales las proyecciones de la variable aleatoria tengan máximas varianzas en direcciones ortogonales. Con esto último se busca la no coorrelación entre las componentes.

Sin embargo, cuando se trata de proporciones, no es una ventaja encontrar correlaciones nulas, puesto que ellas naturalmente están correlacionadas (con tendencia a ser éstas negativas), por su naturaleza.

La segunda característica se presenta con frecuencia en datos composicionales, aunque no en general. Consiste en posibles curvaturas mostradas por conjuntos de estos datos. Como se dijo antes, un simple análisis en componentes principales dá una transformación lineal de los datos, por lo que las coordenadas que se obtienen no pueden adaptarse a las curvaturas eventuales.

Los datos composicionales $X^{(p+1)}$ se suelen tratar de tres distintas maneras:

La primera es hacerlo tal como vienen, como $X^{(p+1)}$, con la restricción que la suma de sus componentes es 1.

El resultado es que la matriz de varianzas-covarianzas es singular, de rango p. Por lo tanto tiene un valor propio cero. Su correspondiente vector propio tiene coordenadas iguales, lo cual es consecuencia de que la suma de las coordenadas de X es 1. Por lo tanto los demás, por ser ortogonales con él, son contrastes o comparaciones lineales. Luego las p componentes principales con varianza mayor que cero, son contrastes de proporciones de un mismo total. La interpretación de las correlaciones nulas resulta difícil., por tratarse precisamente de proporciones.

La segunda forma es desechando una componente, elegida arbitrariamente, transformándose el vector de $X^{(p+1)} \in \mathbb{H}^{p+1}$ a un vector $X_{-j}^{(p)} \in S^p$, en que X_{-j} es el vector X al cual se le ha eliminado la j-ésima coordenada. Sin embargo siempre existe una restricción sobre el vector, por lo que no se soluciona el problema de la interpretación; la restricción, ahora, es que la suma es menor que 1. Además el resultado será dependiente de la elección que se haga de la componente a eliminar.

La tercera forma de tratar datos composicionales es mediante la transformación

$$X \rightsquigarrow \frac{X_{-j}}{x_j} \in \mathbb{R}^p,$$

en que se ha eliminado la coordenada x_j, escogida arbitrariamente, y se ha dividido las demás por ella. El resultado también dependerá de la variable eliminada, y la transformación mantiene la linealidad que no permite manejar las curvaturas adecuadamente.

1.1. LA TRANSFORMACION LOG - RAZON.

La transformación propuesta por J. Aitchison para el tratamiento de datos composicionales es

$F : \mathbb{H}^{p+1} \longrightarrow \mathbb{R}^p$ en que

$$F = \begin{bmatrix} f_1(X) \\ f_2(X) \\ \vdots \\ f_p(X) \end{bmatrix}$$

con $f_i(X) = f_i(x_1, x_2, \ldots, x_{p+1}) = (\log \frac{x_1}{x_j}, \log \frac{x_2}{x_j}, \ldots, \log \frac{x_{j-1}}{x_j}, \log \frac{x_{j+1}}{x_j}, \ldots, \log \frac{x_{(p+1)}}{x_j})$, (36)

Se denotará

$$F(X) = \log\left(\frac{X_{-j}}{x_j}\right)$$

en que $X \in \mathbb{R}^{p+1}$ y X_{-j} es el vector X al que se le ha omitido la coordenada x_j.

La matriz de varianzas-covarianzas de los datos transformados es

$$\Omega_j = \mathrm{Var}(F(X)) = \mathrm{Var}\left(\log \frac{X_{-j}}{x_j}\right) \qquad (37)$$

La no-linealidad de la función logarítmica permite manejar la posible curvatura de los datos. Subsiste el problema de la asimetría, pues distintas elecciones del divisor x_j conducirán a distintas componentes principales.

Sea $A^* = (a_1^*, a_2^*, \ldots, a_p^*)'$ uno de los vectores que definen los ejes principales. La correspondiente componente principañ del vector F es

$$\sum_{i=1}^{p} a_i^* f_i = \sum_{\substack{i=1 \\ i \neq j}}^{p+1} a_i \log\left(\frac{x_i}{x_j}\right) \qquad (38)$$

con

$$a_k = \begin{cases} a^*_k & \text{si } 1 \leq k < j \\ a^*_{k-1} & \text{si } j < k \leq p \end{cases}$$

Si se define $a_j = -\sum_{\substack{i=1 \\ i\neq j}}^{p+1} a_i$, entonces

$$\sum_{i=1}^{p+1} a_i = 0 \tag{39}$$

por lo que $A = (a_1, a_2, \ldots, a_{p+1})'$ es un contraste.

Además

$$\sum_{\substack{i=1 \\ i\neq j}}^{p+1} a_i \log \frac{x_i}{x_j} = \sum_i a_i (\log x_i - \log x_j) =$$

$$= \sum_i a_i \log x_i - \sum_i a_i \log x_j = \sum_{\substack{i=1 \\ i\neq j}}^{p+1} a_i \log x_i + a_j \log x_j =$$

$$= \sum_{i=1}^{p+1} a_i \log x_i$$

luego

$$\sum_{\substack{i=1 \\ i\neq j}}^{p+1} a_i \log\left(\frac{x_i}{x_j}\right) = \sum_{i=1}^{p+1} a_i \log x_i \tag{40}$$

O sea, los componentes principales en este caso pueden ser expresadas simétricamente en términos de contrastes log-lineales de las p+1 proporciones, cualquiera sea la elección del divisor x_j. En todo caso, debe recordarse que los a_i dependen del x_j que se ha escogido.

Además, como por (39) $\sum_{i=1}^{p+1} a_i = 0$, entonces

$$\left(\sum_{i=1}^{p+1} a_i \right) \left(\sum_{j=1}^{p+1} \frac{\log x_j}{p+1} \right) = 0$$

y

$$\sum_{i=1}^{p+1} a_i \log x_i = \sum_{i=1}^{p+1} a_i \log x_i - \sum_{i=1}^{p+1} a_i \sum_{j=1}^{p+1} \frac{\log x_j}{p+1} =$$

$$= \sum_{i=1}^{p+1} a_i \left(\log x_i - \sum_{j=1}^{p+1} \frac{\log x_j}{p+1} \right) =$$

$$= \sum_{i=1}^{p+1} a_i \left[\log x_i - \log (x_1, x_2, \ldots, x_{p+1})^{\frac{1}{p+1}} \right] =$$

$$= \sum_{i=1}^{p+1} a_i \left[\log \frac{x_i}{(x_1, x_2, \ldots, x_{p+1})^{1/p+1}} \right]$$

Luego

$$\sum_{i=1}^{p} a_i^* f_i = \sum_{i=1}^{p+1} a_i \log x_i = \sum_{i=1}^{p+1} a_i \log \left\{ \frac{x_i}{g(X)} \right\} \qquad (41)$$

con $g\{X\} = (x_1, x_2, \ldots, x_{p+1})^{1/p+1}$, el medio geométrico de las p+1 proporciones x_i.

Esto sugiere que pareciera ser conveniente estudiar los valores y vectores propios de la matriz de varianzas-covarianzas de orden p+1.

$$\Omega = \text{Var} \left[\log \left\{ \frac{X^{(p+1)}}{g(X)} \right\} \right] = (\Omega_{kl}) \qquad (42)$$

con

$$\Omega_{kl} = \text{cov} \left(\log \frac{x_k}{g(X)}, \log \frac{x_1}{g(X)} \right)$$

Ω es semidefinida positiva, por construcción (es matriz de varianzas-covarianzas).

Estudiemos su posible valor propio nulo:

Sea $A = (a_1, a_2, \ldots, a_{p+1})'$ el vector propio asociado a él. Entonces

$$\Omega A = 0$$

$$\sum_{l=1}^{p+1} \Omega_{kl} a_l = 0 \qquad k = 1, 2, \ldots, p+1$$

$$\sum_{l=1}^{p+1} \operatorname{cov}\left(\log \frac{x_k}{g(X)}, \log \frac{x_l}{g(X)}\right) a_l = 0 \tag{43}$$

pero

$$\sum_{l=1}^{p+1} \operatorname{cov}\left(\log \frac{x_k}{g(X)}, \log \frac{x_l}{g(X)}\right) a_l =$$

$$= \operatorname{cov}\left(\log \frac{x_k}{g(X)}, \sum_{l=1}^{p+1} a_l \log \frac{x_l}{g(X)}\right) =$$

$$= \operatorname{cov}\left(\log \frac{x_k}{g(X)}, \log \frac{x_1^{a1}, x_2^{a2}, \ldots, x_{p+1}^{ap+1}}{g(X)^{p+1}}\right) \tag{44}$$

Entonces si $a_1 = a_2 = \ldots = a_{p+1} = 1$, queda

$$\operatorname{cov}\left(\log \frac{x_k}{g(X)}, \log \frac{g(X)^{p+1}}{g(X)^{p+1}}\right) =$$

$$\operatorname{cov}\left(\log \frac{x_k}{g(X)}, \log 1\right) = \operatorname{Cov}\left(\log \frac{x_k}{g(X)}, 0\right) = 0 \tag{45}$$

Luego $\mathbb{1}_{p+1} = (1, 1, \ldots, 1)'$ es un vector asociado al valor propio cero.

Entonces se buscará un máximo de p soluciones de

$$\Omega A = \lambda A \quad \text{o bien}$$

$$(\Omega - \lambda I_{p+1}) A = 0 \tag{46}$$

con $A_1, A_2, \ldots, A_p$ vectores de norma 1.

Al ser ortogonales con el vector propio $\mathbb{1}$, son todos contrastes log-lineales de las proporciones x_i, como se quería.

Si consideramos la matriz

$$B_j = \begin{bmatrix} I_{j-1} & -\mathbb{1}_{j-1} & 0 \\ 0 & -\mathbb{1}_{d-j+1} & I_{d-j+1} \end{bmatrix}_{p \times (p+1)} \tag{47}$$

$$B_j = \begin{bmatrix} \overbrace{1 \quad \cdots \quad -1}^{j-1} & \overbrace{\quad}^{d-j+1} \\ \ddots \quad -1 & \\ 1 \quad -1 \; 1 & 1 \quad \cdots \\ \vdots & \ddots \; 1 \\ -1 & \end{bmatrix}$$

Se establece la relación entre la matriz de varianza-covarianza Ω_j (37) y la matriz Ω de (46),

$$\Omega_j = B_j \, \Omega \, B'_j \tag{48}$$

Premultiplicando (46) por Bj,

$$(B_j \Omega - \lambda B_j) A = 0$$

y observando que

$$A = B_j' A_{-j}, \text{ en que } A_{-j} \text{ es } A_j$$

al que se le ha eliminado la coordenada j-ésima, queda

$$(B_j \Omega B_j' - \lambda B_j B_j') A_{-j} = 0$$

$$(\Omega_j - \lambda H_p) A_{-j} = 0$$

o en general,

$$(\Omega_j - \mu H_p) E = 0 \qquad (49)$$

con μ y E incógnitas y con

$$H_p = B_j B_j' = \begin{bmatrix} 1 & 2 & \cdots & 2 \\ 2 & 1 & \cdots & 2 \\ \vdots & \vdots & \ddots & \\ 2 & 2 & & 1 \end{bmatrix}$$

Entonces el problema simétrico (46) y el problema asimétrico (49) producen componentes principales log-lineales relacionadas por las ecuaciones.

$$\left.\begin{aligned} A &= B_j' E \\ E &= A_{-j} \end{aligned}\right\} \qquad (50)$$

Los vectores propios de la versión asimétrica $E_1, E_2, \ldots, E_p$ no son ortogonales, pero satisfacen las relaciones

$$E_i' H_d E_j = \begin{cases} 1 & \text{si } i=j \\ 0 & \text{si } i \neq j \end{cases} \qquad (51)$$

12. TECNICAS DE ANALISIS CON LA TRANSFORMACION LOG-RAZON.

La ecuación (49) proporciona un análisis en componentes principales que difiere del usual, en que sus ejes no son ortogonales, sino que responden a las relaciones definidas por (51). Tiene las ventajas sobre los procedimientos utilizados usualmente, y que se describieron anteriormente, de que no es lineal por lo que se puede adaptar a las posibles curvaturas, y de que los datos no están restringidos, sino que se mueven en $\mathbb{R}^{p+1}$. Aún subsiste, eso si, el problema de la asimetría, pues el resultado dependerá de la coordenada que se elimine.

Pero se relaciona con el análisis proporcionado por (46), a través de la ecuación (48), o de las relaciones entre vectores propios (50). Estos indican que los vectores propios del problema asimétrico (49) no son sino las proyecciones de los vectores propios del problema simétrico (46), sobre el subespacio que resulta de eliminar la respectiva coordenada que se seleccionó. El problema simétrico (46) tiene las ventajas que señalamos sobre el problema asimétrico, pero además es indiferente con respecto a las coordenadas, sin que deba privilegiarse ninguna de ellas, haciendo el resultado dependiente de cuál se ha escogido. Es todo esto lo que hace atractiva la utilización de la transformación log-razón con un conjunto de datos composicionales.

El procedimiento es el siguiente:

Se efectúa la transformación de los datos de acuerdo a (36). Luego se calcula la matriz de varianzas-covarianzas Ω , y se buscan los valores y vectores propios proporcionados por (46). Estos últimos definen los ejes principales. Como en el análisis de componentes principales usual, los valores propios 1_i, corresponden a los valores de las varianzas de las proyecciones respectivas. La proporción de varianza explicada por las primeras k componentes, también es

$$\frac{1_1 + 1_2 + \ldots + 1_k}{1_1 = 1_2 + \ldots + 1_{p+1}}$$

BIBLIOGRAFIA

1) J. AITCHISON, S.M. SHEN. "Logistic-normal Distributions. Some Properties and Uses". Biometrika, Vol. 67, N° 2, 1980.

2) J. AITCHISON. "A New Approach to Null Correlations of Proportions". Mathematical Geology, Vol. 13, N° 2, 1981.

3) J AITCHISON. "The Statistical Analysis of Compositional Date". J.R. Statistical Society B, Vol. 44, N° 2, 1982.

4) J. AITCHISON. "Principal Component Analysis of Compositonal Data". Biometrika, Vol. 70 N° 1, 1983.

5) R. GNANADESIKAN. "Methods for Statistical Data Analysis of Multivariate Observations". Ed. John Wiley, 1977.

6) P. GREEN, J.D. CARROL. "Mathematical Tools for Applied Multivariate Analysis". Ed. Academic Press, 1976.

7) K.V. MARDIA, J.T. KENT, J.M. BIBBY. "Multivariate Analysis". Ed. Academic Press, 1979.