CAPÍTULO 4

ANÁLISIS DE COMPONENTES PRINCIPALES

4.1 Introducción

Al investigar un fenómeno desconocido se pretende abordarlo con muestras de diferentes variables, en las cuales muchas veces existe una fuerte correlación; las relaciones se pueden interpretar como una medida del fenómeno bajo distintos puntos de vista. En un proceso estadístico que cuenta con un gran número de variables es difícil visualizar sus conexiones, al considerar muchas variables tendremos un número mayor de combinaciones representando los coeficientes de correlación.

Es importante reducir el número de variables para desechar información redundante y optimizar el proceso. El Análisis de Componentes Principales (ACP) propone la transformación a un nuevo conjunto sintético de variables (los componentes principales), que no están correlacionados y se encuentran ordenados de tal forma que los primeros conservan la mayor parte de la variación presente en todas las variables originales.

La técnica de ACP fue desarrollada por Pearson (1901) para luego ser retomada por Hotelling (1933) y posteriormente ser implementadas con el impulso de las computadoras. En éste capítulo presentamos el desarrollo de los componentes principales y su aplicación en el reconocimiento de imágenes.

4.2 Componentes Principales

El análisis de componentes principales (ACP) es una técnica estadística multivariante de simplificación, que permite transformar un conjunto de variables originales correlacionadas entre sí, en un conjunto sintético de variables no correlacionados denominados factores o componentes principales. En esta transformación no se establecen jerarquías entre variables y se elimina la información repetida (Jolliffe, 1986). Las nuevas variables son combinaciones linealmente independientes de las variables originales, ordenadas de acuerdo a la representación de dispersión respecto a la nube total de información recogida en las muestras.

Consideremos un grupo de variables $(x_1, x_2, \dots, x_p)$, cada uno con el mismo número de muestras; cada variable registra información en un vector del mismo tamaño el cual, en forma matricial lo podemos representar de la siguiente manera:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \text{ 100\% de información}$$

n =individuos *p*=variables

La matriz X constituye la tabla de datos u observaciones, el arreglo de columnas y renglones en conjunto representan el 100% de la información.

La comparación de dos individuos *i* y *j* es evaluada con la distancia euclidiana clásica entre:

$$d^2(i,j) = \sum_{P=1}^{p} (x_{iP} - x_{jP})^2 \qquad (4.1).$$

A partir de esta matriz calculamos un nuevo conjunto $(C_1, C_2, \dots, C_p)$ no correlacionados entre si, en la cual sus varianzas vayan decreciendo progresivamente. Cada elemento C_j (j=1,…, p) representa una combinación lineal de las variables originales $(x_1, x_2, \dots, x_p)$.

$$C_j = a_{j1}x_1 + a_{j2}x_2 + \ \dots + a_{jp}x_p$$

$$C_j = a_j'x,$$

donde

$$a_j' = (a_{1j}, a_{2j}, \dots, a_{pj})$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix},$$

a_j' es una matriz de constantes y representa el peso de las variables en cada componente; las componentes explican una parte de la varianza total, pretendiendo encontrar *k* componentes que representen casi toda la varianza de la nube de información.

Condiciones para obtener el primer componente principal:

1. -Para mantener la ortonormalidad de la transformación se impone que:

$$a_j' a_j = \sum_{k=1}^{p} a_{kj}^2 = 1 \qquad (4.1).$$

2. .-La varianza en el primer componente principal C1 sea máxima en la nube de datos.

El primer componente principal representa la mayor varianza observada sujeta a cumplir la condición de ortonormalidad; el segundo componente principal se obtiene calculando los valores de a_2 de tal manera que C_1y C_2 sean no correlacionados (ortogonales). Del mismo modo se eligen los siguientes componentes cada uno con menor varianza que el anterior.

4.3 Obtención Analítica del Primer Componente Principal

Consideremos una matriz de datos en desviaciones respecto a su media *S*; el primer componente principal viene dado por:

$$\begin{bmatrix} C_{11} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} a_{11} \\ \vdots \\ a_{1p} \end{bmatrix} \leftrightarrow C_1 = X a_1 \qquad (4.2),$$

donde C_1 es una variable estandarizada con respecto a su media.

Verificamos que:

$$Var(C_1) = \frac{1}{n}\sum C_{1i}^2 = \frac{1}{n} C_1' C_1 = \frac{1}{n} a_1' X' X a_1$$

$$Var(C_1) = a_1'[\frac{1}{n} X' X] a_1$$

$$\frac{1}{n} X' X = S$$

$$Var(C_1) = a_1' S a_1 \ (máxima\ varianza)$$

$$\sum_{j=1}^{p} a_{1j}^2 = a_1' a_1 = 1 \quad (ortonormalidad)$$

Utilizamos los multiplicadores de Lagrange para maximizar la función sujeta a las condiciones anteriores, donde a_1 es la incógnita que representa el vector con la combinación lineal óptima. Buscamos uno de los valores extremos (para nuestro caso el máximo).

$$L = a_1' S a_1 - \lambda(a_1' a_1 - 1) \quad (4.3),$$

Derivando 4.3 respecto a a_1 e igualando a cero tenemos:

$$\frac{\partial L}{\partial a_1} = 2 S a_1 - 2\lambda a_1 = 0$$

$$\left(S - \lambda I_p\right) a_1 = 0$$

$$S a_1 = \lambda I_p a_1 \qquad (4.4).$$

Resolviendo obtenemos *p* raíces características, donde I_p representa la matriz identidad; para que el sistema tenga una solución distinta de cero la matriz $\left(S - \lambda I_p\right)$ tiene que ser singular, es decir, que su determinante debe ser igual a cero $\left|\left(S - \lambda I_p\right)\right| = 0$. Tomando el máximo de los autovalores λ_1 se obtiene el vector característico asociado a_1, este vector brinda la mejor combinación lineal de las variables originales representando la mayor varianza; C_1 lo podemos

asociar al eje principal de la nube de datos en un nuevo sistema de referencia, es decir, que sobre C_1se presenta la mayor distribución de los datos.

Se cumple que:

$$Var(C_1) = Xa_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1 = a_1' \lambda_1 I_p a_1 = \lambda_1 a_1' a_1 = \lambda_1 1 = \lambda_1.$$

Para obtener la segunda componente realizamos lo siguiente:

$$C_2 = Xa_2$$

$$Var(C_2) = a_2' S a_2.$$

Condiciones

$$a_2' a_2 = 1$$

$$a_2' a_1 = 0 \text{ (ortogonalidad).}$$

Maximizamos la función sujeta a las condiciones con los operadores de Lagrange:

$$L = a_2' S a_2 - \lambda(a_2' a_2 - 1) - m a_2' a_1,$$

derivando e igualando a cero para encontrar el máximo tenemos:

$$\frac{\partial L}{\partial a_2} = 2Sa_2 - 2\lambda a_2 - m a_1 = 0 \qquad (4.5)$$

Multiplicando por a_1

$$2a_1 S a_2 - m = 0$$

$$2a_1 S a_2 = m$$

Como $a_2' S a_1$ no expresa ninguna relación entre las variables entonces:

$$a_2' S a_1 = 0,$$

por los tanto

$$m = 0.$$

De este modo y retomando la ecuación 4.5 tenemos:

$$(S - \lambda I_p)a_2 = 0.$$

Aplicando el mismo razonamiento que el primer componente principal, los coeficientes a_2 del segundo componente principal C_2, corresponde al segundo mayor autovalor λ_2;de igual forma se procede para obtener los demás componentes; cada uno de los componentes se puede representar como el producto de una matriz formada por los autovectores multiplicada por el vector *X* que contiene las variables originales, de tal forma, que la matriz de covarianzas del nuevo conjunto optimizado tendrá valores diferentes de cero solo en su diagonal:

$$\Lambda = \begin{bmatrix} \lambda_1 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_p \end{bmatrix}.$$

Lo cual es de esperarse ya que $C_1, C_2, \dots, C_p$ no están correlacionadas formando un conjunto ortogonal.

Tenemos que Λ

$$\Lambda = \mathrm{Var(C)} = a'\mathrm{Var(X)}a = a'Sa$$

Por propiedades

$$S = a\Lambda a'.$$

Esta expresión simplifica y reduce el número de operaciones a realizar.

Por propiedades matriciales se reduce drásticamente el número de autovalores; por citar un ejemplo, de un conjunto de 40 imágenes cada una con 10 000 pixeles, el conjunto de datos se puede representar con 10 000 autovalores correspondientes a la cantidad de individuos, o de igual

manera el mismo conjunto se puede describir con 40 autovalores que describen la cantidad del variables.

Cada auto valorλ_i correponde a la varianza de su respectiva componente C_i, sumando todos los autovalores obtenemos la varianza total de las componentes.

$$\sum_{i=1}^{p} Var(C_i) = \sum_{i=1}^{p} \lambda_i = Traza(\Lambda)$$

$$Traza(\Lambda) = Traza(a'Sa) = Traza(\mathrm{S}) = \sum_{i=1}^{p} Var(X).$$

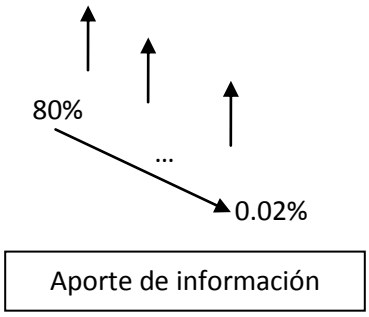
Esto demuestra que la suma de las varianzas de las variables originales y la suma de las varianzas de los componentes principales son idénticas. La conservación de la varianza permite cuantificar en porcentaje la dispersión que recoge cada componente principal y del mismo modo, cuantificar el porcentaje acumulado de los primeros *k* componentes.

$$\frac{\sum_{i=1}^{k} \lambda_i}{Traza(\Lambda)} \quad (4.6),$$

donde se pretende que *k*<<*p.*

Finalmente llegamos a la representación de un nuevo conjunto de variables sintético; cada componente constituye un porcentaje de la información total de la nube de información:

$$X = \begin{bmatrix} C_{11} & \cdots & C_{1p} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{np} \end{bmatrix}$$



Conociendo el aporte de cada componente principal, podemos decidir el número *K* de componentes a utilizar, que representen un porcentaje definido menor al 100%, reduciendo significativamente la cantidad de variables.

En el capítulo 5 observamos que para una prueba con 40 variables originales, más del 90% de la información es concentrada en los primeros 20 componentes principales reduciendo el proceso de computo significativamente.

El coeficiente de correlación r_{ij} entre un componente C_j y una variable X_i, se calcula multiplicando el peso de la variable en ese componente por la raíz cuadrada de su valor propio (Jolliffe, 1986).

$$r_{ij} = \sqrt{\lambda_j}\, a_{ij}$$

i=1,…,*p*
j=1,…,*k*

Estos coeficientes se denominan pesos o cargas factoriales representando la varianza respecto a cada variable.

$$x_1 = r_{11}C_1 + r_{12}C_2 + \ \ldots + r_{1k}C_k$$

$$x_p = r_{p1}C_1 + r_{p2}C_2 + \ \ldots + r_{pk}C_k$$

La matriz de cargas factoriales es de tamaño *pxk,* la suma horizontal del cuadrado de las cargas factoriales es parte de la dispersión total de la variable expuesta por el conjunto de *k* componentes, además, la suma de todos los factores contenidos en la matriz coindice con el porcentaje de dispersión definido por las k componentes, ajustando de igual manera, con la suma de los primeros *k* valores propios elegidos de un total de *p* componentes principales.

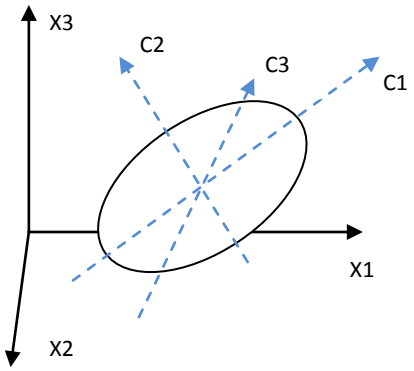


Figura 4.1 Representación gráfica de un conjunto de tres variables, la elipse representa la nube de información.

Los componentes principales estarían representados por los ejes principales del elipsoide (si fuesen tres variables) y el elipsoide conformaría la nube de observaciones.

La magnitud de los eigenvalores corresponde a la varianza de los datos a lo largo de la dirección de su eigenvector correspondiente.

4.4 Reconocimiento de Imágenes Utilizando ACP

Además del propósito de ACP de reducir la dimensiónalidad de los datos, otro de sus objetivos se enfoca a la predicción o reconocimiento de patrones; el método tiene grandes aplicaciones en la identificación de rostros y existe una gama de publicaciones en éste contexto.

Podemos utilizar el mismo concepto y enfocarlo al reconocimiento de espectros de potencia wavelet; cada matriz que genera una imagen la reordenamos como un vector, tomamos sus eigenvalores y formamos un eigenespacio; el eigenespacio es calculado identificando los eigenvectores de la matriz de covarianzas derivado del grupo de espectros que consideramos comparten características similares *precursoras o no precursoras.*

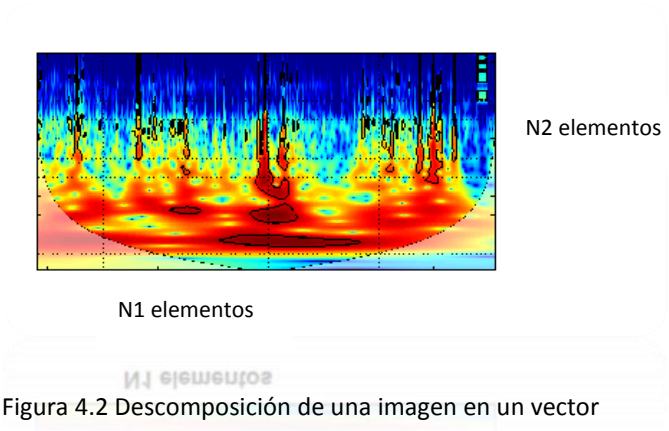


Figura 4.2 Descomposición de una imagen en un vector

Supongamos que Γ es un vector de tamaño Nx1 correspondiente al espectro de potencia wavelet; el propósito es conducirlo a un espacio de menor dimensión, de éste modo un conjunto de M espectros lo podemos representar como una matriz Γ_{ij}dónde el número de columnas son los M espectros de entrenamiento y el número de renglones son los elementos que conforman la imagen.

- Obtenemos el promedio del grupo de espectros

$$\Psi = \frac{1}{M}\sum_{i=1}^{M}\Gamma_{\mathrm{i}}$$

- Sustraemos el espectro promedio de cada elemento del grupo

$$\Phi_{\mathrm{i}} = \Gamma_{\mathrm{i}} - \Psi$$

- Obtenemos la matriz de covarianzas

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_{\mathrm{n}}\Phi_{\mathrm{n}}^{\mathrm{T}} = AA^{T}$$

La matriz A es de tamaño NxM, observamos que la matriz *C* se encuentra definida con un gran número de elementos al resultar de tamaño NxN, donde N representa el total de elementos en cada espectro de potencia wavelet.

$$A = [\Phi_1, \Phi_2, \dots, \Phi_M]$$

Podemos considerar la multiplicación $A^T A$ que comparte los mismos eigenvectores υ_i que AA^T reduciendo significativamente el número de operaciones (Turk y Pentland, 1991).

$$A^T A v_i = \mu_i \upsilon_i$$

Pre multiplicando por A

$$AA^T A v_i = \mu_i A \upsilon_i,$$

sí $\quad C = AA^T \;\; y \;\; u_i = A v_i$

$$C u_i = \mu_i u_i$$

Cada espectro Φ_i en el conjunto Γ_{ij} puede ser representada como una combinación lineal de los mejores *K* eigenvectores.

$$\Phi_i^{apox} - \Psi = \sum_{i=1}^{K} u_i^T \Phi_i u_i.$$

Donde u_i son los eigenvectores, cada espectro Φ_i es representado en esta base como un vector:

$$\Omega_i = \begin{bmatrix} w_1^i \\ w_2^i \\ \dots \\ w_k^i \end{bmatrix} \; donde \; i = 1,2, \dots, M.$$

Con la base generada podemos hacer el reconocimiento de espectros utilizando los eigenvectores; en base al mínimo error εk (con un umbral definido) podemos determinar si Φ_i pertenece o se aproxima al conjunto de espectros originales; el método simplemente determina qué clase de espectros proporcionan una mejor descripción a un espectro de entrada y encuentra los *k* espectros que minimizan la distancia Euclidiana:

$$\varepsilon k = ||\Omega - \Omega_k||$$

Donde Ω describe la contribución de cada eigenvector en representación de la imagen proyectada, en el caso de Ω_k representa la base con *k* elementos de entrenamiento.

La aplicación y los resultados de esta metodología son presentados en el siguiente capítulo, y es utilizada para corroborar las observaciones realizadas en los espectros de potencia wavelet.