CAPÍTULO III. MÉTODO DE COMPONENTES PRINCIPALES

III.1. Justificación del Método

Cualquier persona dedicada a la investigación o a la estadística, que se enfrente con un conjunto de datos multivariados, podría sentirse abrumada por la gran cantidad de números que éste contiene. El problema resulta aparente, ya que esta gran cantidad de números se encuentran relacionados entre sí, y es necesario estudiar la dependencia que existe no solo entre las variables, sino entre los individuos de tal población o conjunto. Estas relaciones son consecuencia de haber efectuado las mediciones de todas las variables en estudio para cada individuo. La inclinación natural del investigador cuando se encuentre de cara a estos números, será analizarlos con la espera de detectar patrones o características en ellos. Lo anterior implica que cada número en la matriz de datos debe ser juzgado en relación con los otros valores tanto en el mismo renglón como en la misma columna de la matriz, por lo que al realizar una simple inspección visual de la matriz es poco probable el hallar patrones que pudieran existir, empeorando la situación conforme se aumenten ya sea las variables o los individuos. Es posible llegar a encontrar patrones en los datos, cuando sea emplea un método adecuado para analizarlos.

III.2 Conceptos Fundamentales

III.2.1 Conceptos geométricos básicos para entender el modelo

Considérese la situación que se presenta en la figura 3.1 donde se encuentran graficados los puntos P y Q. Si se considera a P como el primer punto, se pueden considerar sus

coordenadas sobre los ejes X_1 y X_2 como x_{11} y x_{12} respectivamente. El primer sufijo de la x en cada caso, proporciona el número ordinal del punto referido, mientras que el segundo hace referencia al eje. Los puntos P y Q entonces pueden representarse por los pares (x_{11},x_{12}) y (x_{21},x_{22}), que dan las coordenadas respectivas a los ejes seleccionados.

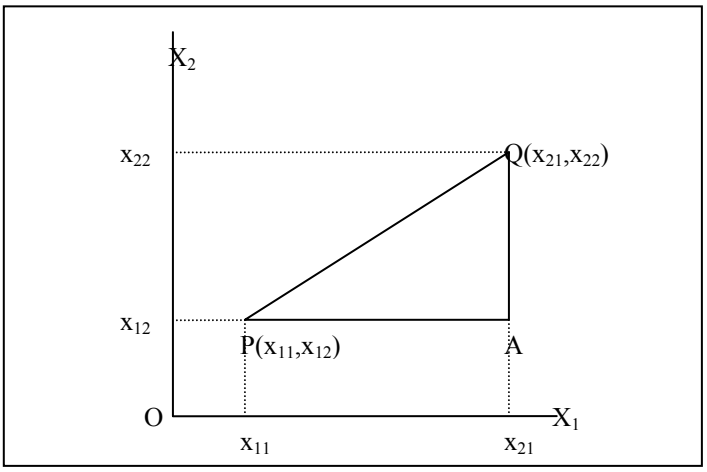

Figura 3.1 Distancia entre dos puntos
Fuente: Krzanowski

La distancia d entre P y Q es calculada fácilmente por aplicación del Teorema de Pitágoras, mediante el cual $PQ^2 = PA^2 + AQ^2$. Al analizar la figura tenemos lo siguiente:

$$PQ^2 = (x_{21} - x_{11})^2 + (x_{22} + x_{12})^2$$

Por tanto, al utilizar notación de sumatoria, es posible escribir:

$$d^2 = \sum_{s=1}^{2} (x_{2s} - x_{1s})^2 \quad ó \quad d = \left\{ \sum_{s=1}^{2} (x_{2s} - x_{1s})^2 \right\}^{\frac{1}{2}}$$

Utilizando un razonamiento similar, si hubiera n puntos en el diagrama entonces la distancia entre cualquiera de ellos, sea el i-ésimo y j-ésimo, sería:

$$d_{ij} = \left\{ \sum_{s=1}^{2} (x_{is} - x_{js})^2 \right\}^{\frac{1}{2}}$$

Matemáticamente, los puntos P y Q son conocidos como vectores que tienen como elementos x_{11}, x_{12} y x_{21}, x_{22} respectivamente. Los vectores son cantidades que poseen longitud y dirección. Por tanto una representación alternativa de estos puntos es como 2 segmentos de línea OP y OQ.

Para encontrar el ángulo entre dos vectores, es necesario aplicar la regla de cosenos de geometría básica. Considerando los puntos de la Fig. 3.2 y sea α el ángulo entre ellos.

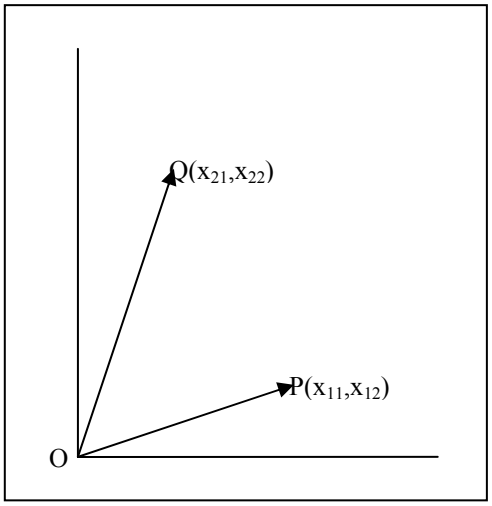

Figura 3.2 Representación Vectorial
Fuente: Krzanowski

$$|QP|^2 = |OQ|^2 + |OP|^2 - 2|OQ| \cdot |OP| \cdot \cos\alpha$$

Al utilizar las fórmulas anteriores para la distancia entre puntos, y la longitud de un vector

$$(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 = x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 2|OQ| \cdot |OP| \cdot \cos\alpha$$

Al desarrollar y posteriormente eliminar términos semejantes,

$$-2x_{11}x_{21} - 2x_{12}x_{22} = -2|OQ| \cdot |OP| \cdot \cos\alpha$$

$$\cos\alpha = \frac{x_{11}x_{21} + x_{12}x_{22}}{|OQ||OP|} = \frac{\sum_{s=1}^{2} x_{1s}x_{2s}}{|OQ||OP|}$$

III.2.2 Aumento de dimensiones

Los conceptos previos se presentaron en términos de dos dimensiones. Si ahora se deseara construir una representación en tres dimensiones, simplemente es necesario añadir un tercer eje X_3 perpendicular a X_1 y X_2, y asignarle coordenadas en la tercera dimensión. Consecuentemente los puntos P y Q tendrían ahora las coordenadas (x_{11},x_{12},x_{13}) y (x_{21},x_{22},x_{23}). Por simple extensión de las reglas anteriores es posible deducir que la distancia d entre P y Q se da por

$$d = \left\{ \sum_{s=1}^{3} (x_{2s} - x_{1s})^2 \right\}^{\frac{1}{2}}$$

y el ángulo α entre vectores

$$\cos\alpha = \frac{x_{11}x_{21} + x_{12}x_{22} + x_{13}x_{23}}{|OQ||OP|} = \frac{\sum_{s=1}^{3} x_{1s}x_{2s}}{|OQ||OP|}$$

donde $|OP|^2 = \sum_{s=1}^{3} x_{1s}x_{2s}$.

Ahora es posible extender todos estos conceptos y reglas para el caso en el que cada punto tenga un número arbitrario de coordenadas. Si se especifican p ejes y coordenadas, y se usa una generalización de las reglas anteriores aplicadas en dos o tres dimensiones, entonces se estaría manipulando lo que los matemáticos conocen como *espacio Euclidiano p-dimensional*. Ahora los puntos P y Q tendrían coordenadas $(x_{11},x_{12},\ldots,x_{1p})$ y

$(x_{21},x_{22},\ldots,x_{2p})$. Con longitudes vectoriales $\left(\sum_{s=1}^{p} x_{1s}^2\right)^{\frac{1}{2}}$ y $\left(\sum_{s=1}^{p} x_{2s}^2\right)^{\frac{1}{2}}$ respectivamente, y con un ángulo α entre ellos dado por

$$\cos\alpha = \frac{\sum_{s=1}^{p} x_{1s}x_{2s}}{\left(\sum_{s=1}^{p} x_{1s}^2 \cdot \sum_{s=1}^{p} x_{2s}^2\right)^{\frac{1}{2}}}$$

La distancia entre los puntos es

$$d = \left\{\sum_{s=1}^{p}\left(x_{2s} - x_{1s}\right)^2\right\}^{\frac{1}{2}}$$

III.2.3 Rotación de ejes

En general se considera que los ejes de una representación geométrica se encuentran fijos, y permanecerán sin cambios en cualquier representación. Cuando la principal preocupación es encontrar la relación entre conjuntos de puntos, la configuración de los puntos es el primordial objeto de estudio, por tanto el marco de referencia utilizado es de interés secundario. Por lo tanto, algunas veces es útil cambiar tal marco, particularmente si este cambio guía a alguna simplificación subsecuente. Frecuentemente se emplea la rotación de ejes, como se muestra en la siguiente figura.

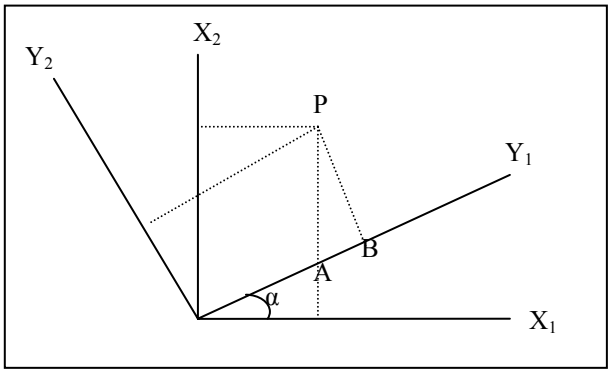

Figura 3.3 Rotación de ejes
Fuente: Krzanowski

Considérese el punto P de la Fig. 3.3, supóngase que cuenta con las coordenadas (x_1,x_2) con referencia a los ejes X_1 y X_2, y posteriormente se rotan los ejes en sentido contrario de las manecillas del reloj con un ángulo α a la posición Y_1 y Y_2. Si las coordenadas de P en los nuevos ejes son y_1 e y_2, entonces tomando como referencia la figura:

$$y_1 = OA + AB$$
$$= \frac{x_1}{\cos\alpha} + y_2 \tan\alpha$$

Despejando x_1

$$x_1 = y_1 \cos\alpha - y_2 sen\alpha \tag{3.1}$$

$$x_2 = PA + AC$$
$$= \frac{y_2}{\cos\alpha} + x_1 \tan\alpha$$

sustituyendo el valor de x_1 en (3.1)

$$x_2 = \frac{y_2}{\cos\alpha} + \frac{sen\alpha}{\cos\alpha}(y_1 \cos\alpha - y_2 sen\alpha)$$

después de desarrollar y hacer sustituciones

$$x_2 = y_1 sen\alpha + y_2 \cos\alpha \tag{3.2}$$

Ahora si (3.1) es multiplicado por cos α, (3.2) multiplicado por sen α, y las ecuaciones resultantes son sumadas, se obtiene

$$x_1 \cos\alpha + x_2 sen\alpha = y_1 \cos^2\alpha + y_1 sen^2\alpha$$

$$y_1 = x_1 \cos\alpha + x_2 sen\alpha \tag{3.3}$$

De manera similar, al multiplicar (3.2) por cos α, (3.1) por sen α, y substraerlas,

$$y_2 = -x_1 sen\alpha + x_2 \cos\alpha \tag{3.4}$$

Por ende si se conocen las coordenadas (x_1,x_2) de un punto P en un par de ejes ortogonales, y se desea deducir sus nuevas coordenadas (y_1,y_2) cuando los ejes se rotan en sentido

contrario de las manecillas del reloj con respecto a un ángulo α, las ecuaciones (3.3) y (3.4) pueden utilizarse para lograr tal fin. En notación matricial se tiene

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\alpha & sen\alpha \\ -sen\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$$

donde

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \qquad \mathbf{A}=\begin{pmatrix} \cos\alpha & sen\alpha \\ -sen\alpha & \cos\alpha \end{pmatrix} \qquad \boldsymbol{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{3.5}$$

Si ahora se supiera que las coordenadas del punto P son (y_1,y_2), y se rotaran los ejes con un ángulo –α, se podrían utilizar las ecuaciones (3.1) y (3.2) para encontrar las coordenadas iniciales. En notación matricial estas ecuaciones se pueden escribir

$$\boldsymbol{x} = \mathbf{B}\boldsymbol{y}$$

$$\boldsymbol{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \mathbf{B}=\begin{pmatrix} \cos\alpha & -sen\alpha \\ sen\alpha & \cos\alpha \end{pmatrix} \qquad \boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

Las matrices **A** y **B** poseen propiedades interesantes. Si se calcula el producto **AB**:

$$\mathbf{AB}=\begin{pmatrix} \cos\alpha & sen\alpha \\ -sen\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\alpha & -sen\alpha \\ sen\alpha & \cos\alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

De manera similar al realizar el producto **BA**, obtenemos el mismo resultado. La matriz que resulta del producto de ambas matrices, es conocida como la matriz identidad, que se denota por **I**. Si se considera la interpretación de **A** y **B** como rotación de ejes, este resultado es perfectamente razonable. Al considerar las coordenadas del punto P. Al rotar los ejes X en un ángulo α, se obtienen $\boldsymbol{y}=\mathbf{A}\boldsymbol{x}$. Ahora al rotar estos nuevos ejes en un ángulo –α se tiene $\mathbf{B}\boldsymbol{y}=\mathbf{BA}\boldsymbol{x}$. Sin embargo, esta es la rotación inversa, la cual regresa a los ejes a su posición original, por tanto estas coordenadas deben ser $\boldsymbol{x}$.

Un segundo aspecto interesante de estas matrices es la relación existente entre sus elementos. Al escribir los renglones de **A** como columnas de una matriz del mismo tamaño, simplemente se obtiene la matriz **B**. Esta relación se expresa al decir que **B** es la transpuesta de **A**. Por tanto, se encuentra de ecuaciones anteriores que **A'A=I**.

Un punto final a notar acerca de las transformaciones que definen los cambios en las coordenadas de P al rotar los ejes con un ángulo α es que las ecuaciones (3.1) a (3.4) pueden escribirse como

$$x_1 = b_{11}y_1 + b_{12}y_2 \tag{3.6}$$

$$x_2 = b_{21}y_1 + b_{22}y_2 \tag{3.7}$$

$$y_1 = a_{11}x_1 + a_{12}x_2 \tag{3.8}$$

$$y_2 = a_{21}x_1 + a_{22}x_2 \tag{3.9}$$

donde los valores a y b son constantes. Por lo cual se puede decir que una rotación de ejes induce una transformación lineal en los valores de las coordenadas de cualquier punto P.

Las ideas anteriores para dos dimensiones generalizan inmediatamente al caso abstracto del espacio Euclidiano p-dimensional. Si las coordenadas del punto P respecto a p ejes mutualmente ortogonales $X_1,X_2,\ldots X_p$ son $(x_1,x_2,\ldots x_p)$, y si estos ejes son rotados de una manera rígida a una nueva posición $Y_1,Y_2,\ldots,Y_p$, entonces las coordenadas $(y_1,y_2,\ldots,y_p)$ de P respecto a los nuevos ejes están dados por la transformación lineal

$$\begin{aligned} y_1 &= a_{11}x_1 + \ldots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + \ldots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + \ldots + a_{pp}x_p \end{aligned}$$

Los distintos valores a son constantes, determinados por la posición precisa de la rotación. Esta transformación lineal puede escribirse de una manera más compacta en forma matricial.

$$\boldsymbol{y}=\mathbf{A}\boldsymbol{x}$$

donde

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \qquad \mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ a_{21} & \cdots & a_{2p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \qquad \boldsymbol{x}=\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

III.3 Técnica de Componentes Principales

Esta metodología tiene su origen con K. Pearson como un medio para ajustar planos mediante mínimos cuadrados ortogonales, pero posteriormente fue propuesto por Hotelling para el propósito particular de analizar estructuras de correlación.

III.3.1 Conceptos geométricos

Considérese la figura 3.4 que exhibe la representación de una muestra de n individuos en el espacio. Supóngase un conjunto de datos imaginarios los cuales exhiben las alturas y pesos de n personas. Esta muestra bidimensional puede ser representada por una gráfica de puntos como la siguiente.

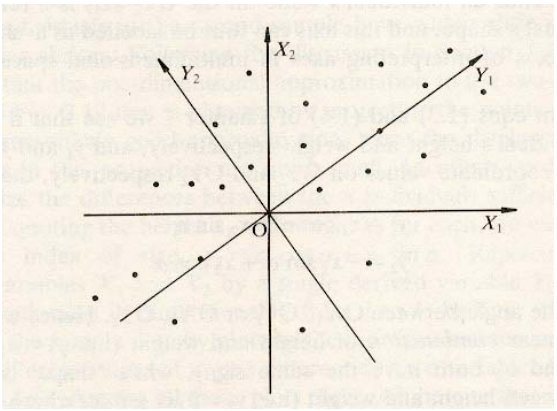

Figura 3.4 Representación de n individuos
Fuente: Krzanowski

Los ejes X_1 y X_2 de esta representación han sido determinados por las variables altura y peso respectivamente. Sin embargo, es la configuración de los puntos lo que interesa realmente, y el marco de referencia con respecto al cual estos puntos son graficados es de cierta manera irrelevante. Por ejemplo, se podrían rotar los ejes a nuevas posiciones Y_1 y Y_2 sin alterar la configuración de puntos, y relacionar los puntos a estos nuevos ejes para un análisis futuro sin cambiar su resultado.

Puede ocurrir que estos nuevos ejes logren conducir a un significado útil para el investigador; de hecho pueden tener mayor relevancia que la configuración original de la que fueron tomados. Considérense los nuevos ejes, e imagínese que todos los datos se comprimen sucesivamente en cada uno de los ejes. Los puntos que se encuentren en la extrema derecha de Y_1 corresponderán a individuos que tienen valores grandes tanto de estatura como de peso; mientras que los puntos a la extrema izquierda corresponderán a individuos con valores pequeños tanto de altura como peso. Por lo tanto el valor de un individuo en el eje Y_1 puede pensarse como un reflejo del tamaño del individuo, y por

tanto este eje puede llamarse eje ‘tamaño’. Ahora considerando el eje Y_2, los puntos hasta arriba de este eje tenderán a corresponder a personas cuyo peso es grande en relación con su altura; mientras que los puntos en la parte inferior tenderán a corresponder a individuos con gran altura en comparación con su peso. Un punto en el medio del eje corresponderá de un individuo cuyo peso y altura están aproximadamente en la proporción correcta. Por lo tanto el valor de un individuo en el eje Y_2 es un reflejo de su forma, y a este eje puede llamársele como eje ‘forma’.

Ahora, si x_1 y x_2 son la altura y peso de un individuo respectivamente, y y_1 e y_2 son las coordenadas de tal individuo en los ejes Y_1 e Y_2, entonces (por resultados de la sección anterior)

$$y_1 = x_1 \cos\alpha + x_2 sen\alpha$$
$$y_2 = -x_1 sen\alpha + x_2 \cos\alpha$$

donde α es el ángulo entre X_1, Y_1 o X_2, Y_2. Por tanto se puede considerar que ‘tamaño’ es una combinación lineal de la altura y el peso, mientras que ‘forma’ es un contraste lineal entre la altura y el peso.

Ahora supóngase que se cuenta con otra configuración de puntos, como se muestra en la figura 3.5 para una muestra distinta. Esta vez, sin embargo, nótese que mientras hay una gran cantidad de valores muestrales en el eje Y_1, la propagación de valores en el eje Y_2 es relativamente pequeña. Esto significa que las personas tienen tamaños muy diferentes pero formas similares. Se podrían caracterizar la diferencias entre los n individuos suficientemente bien si, en lugar de considerar altura y peso para cada uno, simplemente se considerara su índice del tamaño $y_1 = x_1 \cos\alpha + x_2 sen\alpha$. Reemplazando las variables originales X_1 y X_2 por una única variable Y_1 se lleva a cabo una reducción en dimensión de

2 a 1, ya que ahora simplemente se podrían representar los datos muestrales al graficar los valores de cada individuo por su valor Y_1. Valores diferentes de α dan diferentes variables Y_1 y por tanto gráficas diferentes. Entre todas estas gráficas, habrá una considerada como la 'mejor', en el sentido de que provee la impresión más verdadera de la relación que existe entre los n puntos en la figura bidimensional. No resulta difícil encontrar la condición que lleve a tal mejor reducción. La impresión más verdadera de todas las relaciones será provista por el valor de α que de lugar al menor desplazamiento de todos los puntos de su posición original. Dado que los valores de las coordenadas de los puntos en Y_1 son sus proyecciones ortogonales en la línea OY_1 la solución debe estar dada por la línea para la cual estos desplazamientos ortogonales sean los más pequeños. Un punto típico se indica en la figura por Pi y su proyección ortogonal en OY_1 se denota por Pi'. Lo anterior llevó a Pearson a definir la línea OY_1 de ajuste más pequeño a los puntos, a ser aquella obtenida al minimizar $\sum_{i=1}^{n}(P_iP_i')^2$. Es necesario notar que esta es la línea a través de los puntos que minimiza la suma de cuadrados de sus desplazamientos perpendiculares de ésta, en contraste con líneas regresión las cuales minimizan la suma de cuadrados ya sea de desplazamientos horizontales o verticales.

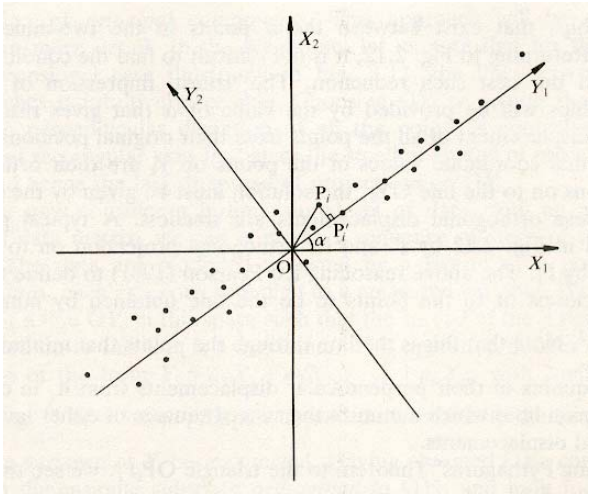

Figura 3.5 Representación de n individuos (distinta muestra)
Fuente: Krzanowski

Aplicando teorema de Pitágoras al triángulo OPiPi', se obtiene

$$(OP_i)^2 = (OP_i')^2 + (P_iP_i')^2$$

Al sumar sobre todos los puntos Pi, se sigue que

$$\sum_{i=1}^{n}(OP_i)^2 = \sum_{i=1}^{n}(OP_i')^2 + \sum_{i=1}^{n}(P_iP_i')^2$$

Por lo tanto

$$\frac{1}{n-1}\sum_{i=1}^{n}(OP_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2 + \frac{1}{n-1}\sum_{i=1}^{n}(P_iP_i')^2$$

al dividir entre (n-1)

De este modo el lado izquierdo de esta ecuación, está fijo para cualquier muestra dada sin importar el sistema de coordenadas que sea empleado. Por tanto escoger OY_1 para minimizar $\frac{1}{n-1}\sum_{i=1}^{n}(P_iP_i')^2$ debe ser equivalente a escoger OY_1 para maximizar $\frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2$. Debido a que O es el centroide de los puntos, $\frac{1}{n-1}\sum_{i=1}^{n}(OP_i')^2$ es la varianza muestral cuando los individuos tienen valores dados por su coordenada Y_1. Por tanto encontrar la línea OY_1 que minimice la suma de desviaciones perpendiculares cuadradas existente entre los puntos y ésta, es exactamente equivalente a encontrar la línea OY_1 tal que las proyecciones de los puntos en esta tenga una varianza máxima. Esta es una aproximación más efectiva operacionalmente, y fue el punto de inicio de Hotelling para su derivación de componentes principales. Al escoger OY_1 para asegurar la desviación perpendicular más pequeña posible de todos los puntos es equivalente a escoger ejes

rectangulares que den la menor propagación de proyecciones en OY_2 y por tanto la mayor propagación de proyecciones en OY_1.

Si se considera ahora un conjunto de datos p-dimensional, con una matriz de datos asociada (nxp), se puede seguir una secuencia de pasos de la forma anterior. Los datos pueden ser modelados de la manera usual por un conjunto de n puntos en p dimensiones, cada uno de los ejes correspondiendo a variable que se esta midiendo. Por tanto se puede buscar una línea OY_1 en este espacio, tal que la propagación de los n puntos cuando se proyecten en esta línea sea máxima. Esta operación define una variable de la forma $Y_1 = a_1X_1+a_2X_2+\ldots+a_pX_p$, con coeficientes a_i que satisfagan $\sum_{i=1}^{p} a_i^{\,2} = 1$ y determinados por el requerimiento de que la varianza de Y_1 sea maximizada. Al obtener OY_1, se considera el subespacio (p-1) dimensional ortogonal a OY_1, y se busca la línea OY_2 en este subespacio tal que la propagación de puntos cuando estos se proyecten en OY_2 sea tan grande como sea posible. Al obtener OY_2, entonces ahora se considera el subespacio (p-2) dimensional ortogonal tanto a OY_1 como a OY_2, y así sucesivamente. Este proceso puede continuarse hasta que se hayan obtenido p líneas mutuamente ortogonales. Cada una de estas líneas define a una variable $Y_i = a_{i1}X_1+a_{i2}X_2+\ldots+a_{ip}X_p$ (i=1,2,…,p), donde las constantes a_{ij} son determinadas por el requerimiento de que la varianza de Y_i sea un máximo pero sujeto a la restricción de ortogonalidad con cada Y_i (j<i). Las Y_i obtenidas son conocidas como las componentes principales del sistema, y el proceso para obtenerlas es llamado análisis de componentes principales.

La utilidad del análisis de componentes principales en datos multivariados debe ser evidente ahora. El modelo geométrico p-dimensional formado de la muestra puede ser considerado como la figura verdadera de los datos. Si se desea obtener la mejor r (< p) representación dimensional de esta verdadera figura p-dimensional, entonces simplemente es necesario proyectar los puntos en el subespacio r-dimensional definido por las primeras r componentes principales Y_1, Y_2,…,Y_r. Esto se hace muy fácilmente como sigue. La matriz original de datos contiene los valores x_{ij} observados en la j-ésima variable del i-ésimo individuo en la muestra (i=1,…,n; j=1,…,p). El análisis de componentes principales proporciona los r conjuntos de coeficientes en las definiciones de los componentes principales $Y_i = a_{i1}X_1+a_{i2}X_2+\ldots+a_{ip}X_p$ (i=1,…r). Los r valores de los componentes principales para el s-ésimo elemento muestral está dado por

$$
\begin{aligned}
y_{s1} &= a_{11}x_{s1} + a_{12}x_{s2} + \ldots + a_{1p}x_{sp} \\
y_{s2} &= a_{21}x_{s1} + a_{22}x_{s2} + \ldots + a_{2p}x_{sp} \\
&\vdots \\
y_{sr} &= a_{r1}x_{s1} + a_{r2}x_{s2} + \ldots + a_{rp}x_{sp}
\end{aligned}
$$

Estos r valores y_{s1}, y_{s2},…,y_{sr} son usualmente llamados los r registros componentes principales para el s-ésimo individuo. Al repetir este cálculo para todos los elementos muestrales se convierte a la matriz de datos original con n renglones y p columnas en una matriz de registros componentes principales con n renglones y r columnas. Estos valores pueden entonces ser graficados en r ejes ortogonales para obtener la proyección requerida de la muestra en el subespacio r-dimensional definido por las primeras r componentes principales. Por supuesto, es caso más útil es aquel cuando r=2, ya que es posible graficar los pares (y_{i1},y_{i2}) en un diagrama de puntos. Tal gráfica proporciona un acercamiento bidimensional a la verdadera configuración de puntos en p dimensiones.

El hecho de poder describir razonablemente ésta como la mejor representación bidimensional para observar los datos esta justificado por la propiedad de la varianza de los componentes principales, y su conexión con una proyección ortogonal. Debido a que las primeras dos componentes principales, son combinaciones lineales de las variables originales, que tienen la varianza muestral más grande asociada con ellas, el plano definido por estas dos componentes es el subespacio bidimensional del espacio original p-dimensional que ha asociado con este la mayor diseminación de todos los n puntos. Por tanto, es el plano en el cual hay la mayor posibilidad de mostrar todas las propiedades del enjambre original de n puntos.

La belleza de este método radica en que si posteriormente se desea encontrar el mejor ajuste de representación tridimensional de puntos, solo se necesita añadir los registros de la tercera componte principal Y_3 a los registros que ya existen de las primeras dos componentes principales Y_1 y Y_2 y usar las triadas (y_{i1}, y_{i2}, y_{i3}) como las coordenadas de los puntos en tres dimensiones.

En general, se supone que un vector (px1) $\mathbf{X}'=(X_1,X_2,\ldots,X_p)$ de valores de variables es observado para cada una de las n unidades independientes, dando lugar a la matriz usual de datos (nxp) en la que su (i,j)-ésimo elemento es x_{ij}. Por medio del vector x_i, se denotan los valores observados en la unidad i-ésima. La media de la j-ésima variable en la muestra esta dada por $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Sean todas estas medias recolectadas de manera conjunta en el vector de la media muestral $\bar{x}' = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)$. La varianza de la j-ésima variable esta dada

por $s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, y la covarianza entre la j-ésima y la k-ésima variable esta dada por $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$.

Sean todas estas varianzas y covarianzas recolectadas conjuntamente en una matriz muestral de covarianza S la cual tiene al (j,k)-ésimo elemento s_{jk}. Al expandir el lado derecho de la matriz como un producto de matrices, se puede verificar que

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

La matriz

$$C = (n-1)S = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

cuyo (j,k)-ésimo elemento $c_{jk} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ es referida frecuentemente como la matriz de suma de cuadrados y productos corregida.

III.3.2 Detalles matemáticos

Considerando una variable escalar Y como una combinación lineal de $a_1X_1+a_2X_2+\ldots+a_pX_p$ de las variables originales X_i. Se puede escribir de manera reducida Y=a'X, donde $a'=(a_1,a_2,\ldots,a_p)$. Entonces el valor de Y correspondiente al i-ésimo miembro muestral estará dado por $y_i=a_1x_{i1}+a_2x_{i2}+\ldots+a_px_{ip}=a'x_i$, mientras la media de Y sobre los n miembros de la muestra claramente será igual a $\bar{y} = a_1\bar{x}_1 + a_2\bar{x}_2 + \ldots + a_p\bar{x}_p = a'\bar{x}$. A continuación,

considérese la varianza muestral de Y, la cual está dada por $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Pero $y_i - \bar{y} = a'x_i - a'\bar{x} = a'(x_i - \bar{x})$. Asimismo, debido a que la transpuesta de un escalar es el mismo escalar, es posible escribir de manera equivalente $(y_i - \bar{y}) = (x_i - \bar{x})'a$. Por tanto, al multiplicar ambos términos conjuntamente, se encuentra lo siguiente

$$(y_i - \bar{y})^2 = a'(x_i - \bar{x})(x_i - \bar{x})'a$$

tal que

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} a'(x_i - \bar{x})(x_i - \bar{x})'a$$
$$= a'\left\{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'\right\}a$$

ya que a es constante para toda i, y por tanto

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = a'\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'\right\}a$$

=a’Sa de la definición de S

Es fácil verificar, al multiplicar vectores, que la forma cuadrática a’Sa se puede expresar en términos de los elementos de a y S como $\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j s_{ij}$.

Habiendo establecido lo anterior puede realizarse la derivación de los componentes principales. En la explicación geométrica explicada con anterioridad, se mostró que la primera componente principal podía ser vista como la línea en el espacio p-dimensional que contiene n puntos con respecto a la cual las proyecciones de los n puntos tienen varianza máxima. Por tanto, algebraicamente, es posible definir la primera componente principal como una combinación lineal Y_1=a’$_1$X de las variables originales lo cual conlleva

al máximo valor de a'Sa sujeto a la restricción $a'_1a_1=1$. Por lo tanto, se requiere encontrar un vector a_1 que satisfaga estas condiciones. Puede verificarse que las condiciones para que el vector a_1 maximice a'_1Sa_1 sujeto a la restricción $a'_1a_1=k$ son precisamente las mismas para que el vector a_1 maximice $a'_1Sa_1-l_1(a'_1a_1-k)$, donde l_1 es una constante conocida como multiplicador de Lagrange. Esta última maximización puede realizarse en el mismo sentido que en cálculo diferencial. En este caso se procede de la siguiente manera.

Sea

$$V_1 = a_1' S a_1 - l_1 (a_1' a_1 - 1)$$
$$= \sum_{i=1}^{p} \sum_{j=1}^{p} a_{1i} a_{1j} s_{ij} - l_1 \left(\sum_{i=1}^{p} a_{1i}^2 - 1 \right)$$

donde $a_1' = (a_{11}, a_{12}, \ldots, a_{1p})$. Entonces

$$\frac{\partial V_1}{\partial a_{1k}} = 2 \sum_{j=1}^{p} s_{kj} a_{1j} - 2 l_1 a_{1k} \quad (k = 1, \ldots, p)$$

Para encontrar el vector a_1' que maximice V_1, se fija $\frac{\partial V_1}{\partial a_{1k}}=0$ para todo k y se resuelve el conjunto resultante de ecuaciones simultáneas.

$$\frac{\partial V_1}{\partial a_{1k}} = 0 \rightarrow \sum_{j=1}^{p} s_{kj} a_{1j} = l_1 a_{1k}$$

El lado izquierdo de esta ecuación es el k-ésimo elemento de Sa_1, mientras que el lado derecho es el k-ésimo elemento de l_1a_1. Por tanto cuando todas las k ecuaciones son tratadas simultáneamente, se sigue que el valor a_1 que maximice debe satisfacer

$$Sa_1 = l_1 a_1 \quad (S - l_1 I) a_1 = 0$$

Este es un conjunto homogéneo de p ecuaciones en p desconocidas, y la teoría de ecuaciones señala que para una solución no trivial se requiere

$$|S-l_1I|=0$$

Por lo tanto l_1 es un eigenvalor de S, y la solución a_1 es su correspondiente eigenvector normalizado tal que $a_1'a_1 = 1$. Sin embargo hay p eigenvalores de S, así que todavía se debe determinar cual de estos es el requerido. Si se multiplica Sa_1 por a_1', se obtendrá $a_1'Sa_1=l_1a_1'a_1$, lo que conlleva a $l_1=a_1'Sa_1$. Dado que l_1 es la varianza muestral de Y_1, y dado que se intenta maximizar esta varianza entonces l_1 debe ser escogido de tal manera que sea el eigenvalor más grande de S. Se sigue por tanto que los coeficientes a_1 en el primer componente principal $Y_1=a_1'$ X están dados por los elementos del eigenvector a_1 que corresponde al eigenvalor más grande de S.

Ahora considérese el segundo componente principal. La explicación geométrica anterior estableció que era la línea, en el espacio p-dimensional, que es ortogonal a la línea que define a la primera componente principal, y que es aquella tal que las proyecciones de los n puntos en esta línea dan como resultado la varianza más grande posible entre todas las líneas. Entonces se busca una segunda combinación lineal $Y_2 =a_2'X$ de las variables originales. Se cumplen las mismas condiciones para esta combinación como antes, $a_2'a_2 =1$. Además, esta línea debe ser ortogonal a la definida por la primer componente principal, condición por la cual $a_2'a_1=a_1'a_2=0$. La varianza de Y_2 es claramente $a_2'Sa_2$, así que la maximización de esta varianza sujeta a las restricciones anteriores de nuevo tendrá que involucrar multiplicadores de Lagrange. Con 2 restricciones son necesarios dos multiplicadores, l_2 y m, y se requiere maximizar

$$V_2 = a_2' S a_2 - l_2(a_2' a_2 - 1) - m(a_2' a_1)$$
$$= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{2i} a_{2j} s_{ij} - l_2\left(\sum_{i=1}^{p} a_{2i}^2 - 1\right) - m\left(\sum_{i=1}^{p} a_{1i} a_{2i}\right)$$

De esta manera

$$\frac{\partial V_2}{\partial a_{2k}} = 2\sum_{j=1}^{p} s_{kj} a_{2j} - 2l_2 a_{2k} - m a_{1k} \quad k = 1,..., p$$

Identificando los términos en el lado derecho como los k-ésimos elementos de $2Sa_2$, $2l_2a_2$ y ma_1 respectivamente, se advierte que al fijar $\frac{\partial V_2}{\partial a_{2k}}$ =0 para toda k lleva a la ecuación

$$(S - l_2 I)a_2 = \frac{1}{2} m a_1$$

Al multiplicar por a_1', se tiene que

$$a_1' S a_2 = \frac{1}{2} m$$

El hecho de multiplicar $Sa_1 = l_1a_1$ por a_2', produce $a_2'Sa_1 = 0$. Dado que $a_1'Sa_2$ es una cantidad escalar y S es una matriz simétrica, entonces $a_1'Sa_2=a_2'Sa_1=0$. Haciendo sustituciones se tiene que m=0, y como resultado que los coeficientes a_2 de la segunda componente principal también satisfacen $(S-l_2I)a_2 = 0$. Un razonamiento enteramente análogo al utilizado para la primera componente, junto con el hecho que la varianza de la segunda componente debe ser máxima después de que la primera componente ha sido considerada, muestra que los coeficientes del segundo componente principal $Y_2=a_2'X$ están dados por los elementos del eigenvector a_2 correspondiente al segundo eigenvalor más grande l_2 de S.

El proceso anterior puede continuarse para todos los componentes principales. De modo que en caso del j-ésimo componente, se busca una combinación lineal $Y_j=a_j'X$ de las variables originales que sea ortogonal a todas las combinaciones previas. La maximización de la varianza sujeta a estas restricciones lleva a una maximización de una expresión que involucra j multiplicadores de Lagrange.