5

COMPONENTES PRINCIPALES

Karl Pearson (1857-1936) *Científico británico. Inventor del contraste que lleva su nombre y uno de los fundadores de la Estadística en el siglo* XIX. *Sus trabajos sobre ajustes ortogonales precedieron el análisis de componentes principales. Fue Catedrático de matemáticas y después de Eugenesia en la Universidad de Londres. Fundador con Weldon, y con el apoyo económico de Galton, de la prestigiosa revista de estadística* Biometrika.

5.1. Introducción

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

El análisis de componentes principales tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20 por 100 de las originales) expliquen la mayor parte (más del 80 por 100 de la variabilidad original).

La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901). Su utilidad es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña observaciones de un espacio general p-dimensional. En este sentido, componentes principales es el primer paso para identificar las posibles variables *latentes*, o no observadas que generan los datos.

2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

En este capítulo presentamos únicamente esta técnica como una herramienta exploratoria. El problema de inferir si las propiedades de reducción de la dimensión encontradas en los datos puede extenderse a la población de la que provienen se estudiara en el Capítulo 12, análisis factorial.

5.2. Planteamiento del problema

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas los elementos. Supondremos en este capítulo que previamente hemos restado a cada variable su media, de manera que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de covarianzas vendrá dada por $1/n\ \mathbf{X}'\mathbf{X}$.

El problema que se desea resolver es encontrar un espacio de dimensión más reducida que represente adecuadamente los datos. Puede abordarse desde tres perspectivas equivalentes.

a) Enfoque descriptivo

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo. Consideremos primero un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos sobre esta recta mantengan, lo más posible, sus posiciones relativas. Para concretar, consideremos el caso de dos dimensiones ($p = 2$). La Figura 5.1 indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un

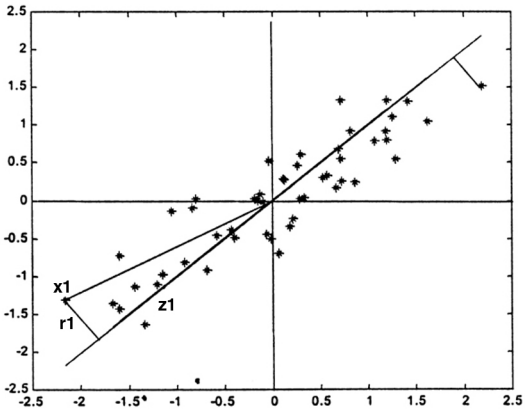


Figura 5.1. Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella.

buen resumen de los datos, ya que la recta pasa cerca de todos los puntos y las distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta. La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si consideramos un punto $\mathbf{x}_i$ y una dirección $\mathbf{a}_1 = (a_{11}, ..., a_{1p})'$, definida por un vector $\mathbf{a}_1$ de norma unidad, la proyección del punto $\mathbf{x}_i$ sobre esta dirección es el escalar:

$$z_i = a_{11}x_{i1} + \ldots + a_{1p}x_{ip} = \mathbf{a}_1'\mathbf{x}_i \tag{5.1}$$

y el vector que representa esta proyección será $z_i\mathbf{a}_1$. Llamando r_i a la distancia entre el punto $\mathbf{x}_i$, y su proyección sobre la dirección $\mathbf{a}_1$, este criterio implica:

$$minimizar \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |\mathbf{x}_i - z_i\mathbf{a}_1|^2 , \tag{5.2}$$

donde $|\mathbf{u}|$ es la norma euclídea o módulo del vector $\mathbf{u}$.

La Figura 5.1 muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipotenusa es la distancia del punto al origen, $(\mathbf{x}_i'\mathbf{x}_i)^{1/2}$, y los catetos la proyeccion del punto sobre la recta (z_i) y la distancia entre el punto y su proyección (r_i). Por el teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i'\mathbf{x}_i = {z_i}^2 + r_i^2, \tag{5.3}$$

y sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2. \tag{5.4}$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^{n} r_i^2$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones z_i son, por (5.1) variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza, y obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados. Este resultado es intuitivo: la recta de la Figura 5.1 parece adecuada porque conserva lo más posible la variabilidad original de los puntos. El lector puede convencerse considerando una dirección de proyección perpendicular a la de la recta en esta figura: los puntos tendrían muy poca variabilidad y perderíamos la información sobre sus distancias en el espacio.

Si en lugar de buscar la dirección que pasa cerca de los puntos buscamos la dirección tal que los puntos proyectados sobre ella conserven lo más posible sus distancias relativas llegamos al mismo criterio. En efecto, si llamamos $d_{ij}^2 = \mathbf{x}_i'\mathbf{x}_j$ a los cuadrados de las distancias originales entre los puntos y $\widehat{d}_{ij}^2 = (z_i - z_j)^2$ a las distancias entre los puntos proyectados sobre una recta, deseamos que

$$D = \sum_i \sum_j (d_{ij}^2 - \widehat{d}_{ij}^2)$$

sea mínima. Como la suma de las distancias originales es fija, minimizar D requiere maximizar $\sum_i \sum_j \widehat{d}_{ij}^2$, las distancias entre los puntos proyectados. Se demuestra en el Apéndice 5.1 que la dirección es la misma que proporciona una variable escalar de varianza máxima.

b) Enfoque estadístico

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, y, se demuestra en el Apéndice 5.2, que la condición para que podamos prever con la mínima pérdida de información los datos observados, es utilizar la variable de máxima variabilidad.

Volviendo a la Figura 5.1 se observa que la variable escalar obtenida al proyectar los puntos sobre la recta sirve para prever bien el conjunto de los datos. La recta indicada en la figura no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino la que minimiza las distancias ortogonales o entre los puntos y la recta y se encuentra entre ambas rectas de regresión.

Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello, calcularemos el plano que mejor aproxima a

los puntos. El problema se reduce a encontrar una nueva dirección definida por un vector unitario, $\mathbf{a_2}$, que, sin pérdida de generalidad, puede tomarse ortogonal a $\mathbf{a_1}$, y que verifique la condición de que la proyección de un punto sobre este eje maximice las distancias entre los puntos proyectados. Estadísticamente esto equivale a encontrar una segunda variable z_2, incorrelada con la anterior, y que tenga varianza máxima. En general, la componente $z_r (r < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p variables originales, con la condición de estar incorrelada con las $z_1, ..., z_{r-1}$ previamente obtenidas.

c) Enfoque geométrico

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. Si consideramos la nube de puntos de la Figura 5.1 vemos que los puntos se sitúan siguiendo una elipse y podemos describirlos por su proyección en la dirección del eje mayor de la elipse. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales, con lo que volvemos al problema que ya hemos resuelto. En varias dimensiones tendremos elipsoides, y la mejor aproximación a los datos es la proporcionada por su proyección sobre el eje mayor del elipsoide. Intuitivamente la mejor aproximación en dos dimensiones es la proyección sobre el plano de los dos ejes mayores del elipsoide y así sucesivamente. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales o incorreladas como veremos a continuación.

5.3. Cálculo de los componentes

5.3.1. Cálculo del primer componente

El *primer componente principal* se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores en este primer componente de los n individuos se representarán por un vector $\mathbf{z}_1$, dado por

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1.$$

Como las variables originales tienen media cero también $\mathbf{z}_1$ tendrá media nula. Su varianza será:

$$\frac{1}{n}\mathbf{z}_1'\mathbf{z}_1 = \frac{1}{n}\mathbf{a}_1'\mathbf{X}'\mathbf{X}\mathbf{a}_1 = \mathbf{a_1'}\mathbf{S}\mathbf{a_1} \tag{5.5}$$

donde $\mathbf{S}$ es la matriz de varianzas y covarianzas de las observaciones. Es obvio que podemos maximizar la varianza sin límite aumentando el módulo del vector $\mathbf{a}_1$. Para que la maximización de (5.5) tenga solución debemos imponer una restricción al módulo del vector $\mathbf{a}_1$, y, sin pérdida de generalidad, impondremos que $\mathbf{a}_1'\mathbf{a}_1 = 1$. Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$M = \mathbf{a_1'}\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a_1'}\mathbf{a_1} - 1)$$

y maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de $\mathbf{a}_1$ e igualando a cero. Entonces

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0,$$

cuya solución es:

$$\mathbf{S}\mathbf{a_1} = \lambda\mathbf{a}_1, \tag{5.6}$$

que implica que $\mathbf{a_1}$ es un vector propio de la matriz $\mathbf{S}$, y λ su correspondiente valor propio. Para determinar qué valor propio de $\mathbf{S}$ es la solución de (5.6), multiplicando por la izquierda por $\mathbf{a}_1'$ esta ecuación,

$$\mathbf{a_1'}\mathbf{S}\mathbf{a_1} = \lambda\mathbf{a_1'}\mathbf{a_1} = \lambda$$

y concluimos, por (5.5), que λ es la varianza de $\mathbf{z_1}$. Como ésta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz $\mathbf{S}$. Su vector asociado, $\mathbf{a_1}$, define los coeficientes de cada variable en el primer componente principal.

Ejemplo 5.1.

Ilustraremos con detalle el cálculo de la primera componente principal con los datos de los logaritmos de las ACCIONES, fichero acciones.dat y Anexo I. Los paquetes estadísticos habituales (Minitab, SPSS, Statgraphics, etc.) proporcionan directamente los componentes principales, pero vamos a indicar los cálculos para el lector interesado.

La matriz de varianzas y covarianzas de estos datos en logaritmos, que ya utilizamos en el Ejemplo 3.5, es,

$$S = \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix}$$

Los valores propios son las raíces de la ecuación:

$$\begin{aligned} &|\mathbf{S} - \lambda\mathbf{I}| = \\ &= \left| \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right| = \\ &= 0.000382 - 0.0628\lambda + 0.64\lambda^2 - \lambda^3 = 0 \end{aligned}$$

Las raíces de este polinomio, obtenidas con Matlas, son $\lambda_1 = 0.521$, $\lambda_2 = 0.113$, $\lambda_3 = = 6.51 \times 10^{-3}$. El vector propio asociado a λ_1 nos da los pesos de la primera componente principal. Para calcularlo, resolvemos el sistema

$$\mathbf{S}\mathbf{a_1} = \lambda_1\mathbf{a}_1$$

que conduce a:

$$\begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix} = 0.521 \times \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix}$$

$$\begin{bmatrix} -0.171a_{11} + 0.15a_{12} - 0.19a_{13} \\ 0.15a_{11} - 0.391a_{12} - 0.03a_{13} \\ -0.19a_{11} - 0.03a_{12} - 0.361a_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

y el sistema es compatible indeterminado. Para encontrar una de las infinitas soluciones tomemos la primera variable como parámetro, x, y resolvamos el sistema en función de x. La solución es,

$$\{a_{11} = x,\ a_{12} = 0.427x,\ a_{13} = -0.562x\}$$

El valor de x se obtiene ahora imponiendo que el vector tenga norma unidad, con lo que resulta:

$$\mathbf{a}_1 = \begin{bmatrix} -0.817 \\ -0.349 \\ 0.459 \end{bmatrix}$$

y el primer componente es

$$Z_1 = -0.817X_1 - 0.349X_2 + 0.459X_3$$

donde X_1, X_2 y X_3 son las variables en logaritmos. Por ejemplo, el valor de esta nueva variable, la primera componente principal, para la primera observación (la primera acción) es

$$z_1 = -0.817 \times \log(3.4) - 0.349 \times \log(89.7) + 0.459 \times \log(30.2) = -1.0049$$

El primer componente principal puede aproximadamente escribirse

$$Z_1 \cong -0.82X_1 + 0.35(X_3 - X_2) + 0.11X_3$$

y utilizando la definición de las variables originales:

$$Z_1 \cong -0.82 \log(d/p) + 0.35 \log(p/d) + 0.11 \log(pN/b)$$

es decir,

$$Z_1 \cong -1.17 \log(d/p) + 0.11 \log(pN/b)$$

que indica que este primer componente depende básicamente de la variable X_1, la rentabilidad por dividendos. Llamando $z_1 = \log Z_1$ este primer componente puede escribirse también como

$$z_1 = \frac{p^{1.27}}{d^{1.16}} \left(\frac{N}{B}\right)^{.09}$$

que es, aproximadamente, de nuevo la variable x_1, el cociente entre el precio de la acción y los dividendos recibidos. Esta variable es la que explica mejor la variabilidad conjunta de las acciones.

Ejemplo 5.2.
Los datos de EPF de la encuesta de presupuestos familiares en España (fichero epf.dat y Anexo I) presentan los gastos medios de las familias españolas para las 51 provincias (Ceuta y Melilla aparecen unidas como una provincia) en nueve epígrafes: X_1 = alimentación, X_2 = vestido y calzado, X_3 = vivienda, X_4 = mobiliario doméstico, X_5 = gastos sanitarios, X_6 = transportes, X_7 = enseñanza y cultura, X_8 = turismo y ocio, X_9 = otros gastos. La matriz de covarianzas resume la variabilidad de estas 9 variables en los 51 elementos observados. Como las distribuciones de los gastos son muy asimétricas, las variables se han expresado en logaritmos. El vector propio asociado al mayor valor propio, 0.348, define la siguiente variable, primer componente principal:

$$\begin{aligned} z_1 = {} & 0.12x_1 + 0.18x_2 + 0.30x_3 + 0.31x_4 + 0.46x_5 + 0.34x_6 \\ & + 0.50x_7 + 0.31x_8 + 0.31x_9 \end{aligned}$$

Se observa que z_1 es una suma ponderada de todos los gastos, con mayor peso, de los gastos en enseñanza y cultura (x_7) y sanitarios (x_5). El menor peso lo tiene el gasto en alimentación (x_1).

Si calculamos los valores de z_1 para las provincias españolas y las ordenamos por esta nueva variable las provincias quedan prácticamente ordenadas por su renta. La primera componente principal tiene, pues, en este caso, una explicación inmediata: redescubre la renta de cada provincia.

5.3.2. Cálculo del segundo componente

Vamos a obtener el mejor plano de proyección de las variables $\mathbf{X}$. Lo calcularemos estableciendo como función objetivo que la suma de las varianzas de $\mathbf{z_1} = \mathbf{X a_1}$ y $\mathbf{z_2} = \mathbf{X a_2}$ sea máxima, donde $\mathbf{a_1}$ y $\mathbf{a_2}$ son los vectores que definen el plano. La función objetivo será:

$$\phi = \mathbf{a_1'} \mathbf{S a_1} + \mathbf{a_2'} \mathbf{S a_2} - \lambda_1(\mathbf{a_1'}\mathbf{a}_1 - 1) - \lambda_2(\mathbf{a_2'}\mathbf{a}_2 - 1) \tag{5.7}$$

que incorpora las restricciones de que las direcciones deben de tener módulo unitario $(\mathbf{a}_i'\mathbf{a}_i) = 1,\ i = 1, 2$. Derivando e igualando a cero:

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0$$

$$\frac{\partial \phi}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 = 0$$

La solución de este sistema es:

$$\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1, \tag{5.8}$$

$$\mathbf{S}\mathbf{a}_2 = \lambda_2 \mathbf{a}_2 \tag{5.9}$$

que indica que $\mathbf{a_1}$ y $\mathbf{a_2}$ deben ser vectores propios de $\mathbf{S}$. Tomando los vectores propios de norma uno y sustituyendo en (5.7), se obtiene que, en el máximo, la función objetivo es

$$\phi = \lambda_1 + \lambda_2 \tag{5.10}$$

es claro que λ_1 y λ_2 deben ser los dos autovalores mayores de la matriz $\mathbf{S}$ y $\mathbf{a}_1$ y $\mathbf{a}_2$ sus correspondientes autovectores. Observemos que la covarianza entre $\mathbf{z}_1$ y $\mathbf{z}_2$, dada por $\mathbf{a}_1'\ \mathbf{S}\ \mathbf{a}_2$ es cero ya que $\mathbf{a}_1'\mathbf{a}_2 = 0$, y las variables $\mathbf{z}_1$ y $\mathbf{z}_2$ estarán incorreladas. Puede demostrarse (véase el Ejercicio 5.7) que si en lugar de maximizar la suma de varianzas, que es la traza de la matriz de covarianzas de la proyección, se maximiza la varianza generalizada (el determinante de la matriz de covarianzas) se obtiene el mismo resultado.

Ejemplo 5.3.
El segundo componente principal para las variables de gastos de la EPF definidas en el Ejemplo 5.2 es el asociado al segundo valor propio mayor, que es 0,032. El vector propio asociado a este valor propio define la nueva variable:

$$\begin{aligned} z_2 &= 0.05x_1 + 0.16x_2 - 0.17x_3 + 0.07x_4 - 0.21x_5 + 0.29x_6 - \\ &\quad - 0.40x_7 - 0.17x_8 + 0.78x_9 = \\ &= (0.05x_1 + 0.16x_2 + 0.07x_4 + 0.29x_6 + 0.78x_9) - \\ &\quad - (0.17x_3 + 0.21x_5 + 0.40x_7 + 0.17x_8) \end{aligned}$$

Esta variable es aproximadamente la diferencia entre dos medias ponderadas de los gastos. La primera, da sobre todo peso a otros gastos (x_9), y transporte (x_6). En otros gastos están incluidas las transferencias fuera de la provincia a miembros de la familia mayores de 14 años que no residan en ella, podemos conjeturar que esta variable separa las provincias que reciben transferencias de las que las envian. Es también significativo que estas provincias tienen altos gastos en transporte. La primera media ponderada puede considerarse un indicador de cómo esta provincia envía recursos a otras. La segunda media da mayor peso a las variables enseñanza y cultura (x_7) y gastos sanitarios (x_5).

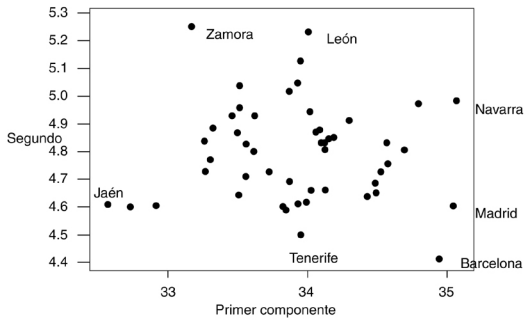


Figura 5.2. Proyección de los datos de la EPF sobre el plano definido por las dos primeras componentes principales.

Este segundo componente va a separar a provincias que envían recursos a otras (alto valor de x_9) y que tienen también altos gastos de transporte, respecto a las que transfieren relativamente poco y tienen altos gastos de educación y sanidad. Las provincias con valores más altos de este componente son Zamora, León, Lugo, Toledo, Huesca, Lérida, Segovia, Soria y Palencia. Estas provincias no han tenido tradicionalmente universidad, por lo que tienen que enviar los estudiantes fuera y tienen bajos costes de educación. Por el contrario, las provincias con valores bajos de este componente z_2 incluyen a Madrid y Barcelona, centros receptores netos de estudiantes de otras provincias, así como a Salamanca, Zaragoza y Tenerife. La Tabla 5.1 presenta la ordenación de las provincias según el primer y segundo componente. La Figura 5.2 representa cada provincia en el plano de las dos primeras componentes principales. Cada punto aparece representado por sus coordenadas respecto a los ejes definidos por las componentes principales y puede interpretarse como la proyección de los puntos, que están en un espacio de dimensión 9, tantos como variables, sobre el plano que mejor mantiene sus distancias relativas, que es el definido por las dos primeras componentes.

Tabla 5.1. Ordenación de las provincias de la EPF, según los dos primeros componentes

Comp. 1	Comp. 2
Navarra	Zamora
Madrid	León
Barcelona	Lugo
Lleida	Toledo
Vizcaya	Huesca
Gerona	Murcia
Baleares	Navarra
Tarragona	Lérida
Guipúzcoa	Segovia
Las Palmas	Soria
⋮	⋮
Ciudad Real	Málaga
Cuenca	Salamanca
Ávila	Cádiz
Teruel	Madrid
Castellón	Badajoz
Orense	Jaén
Zamora	Ceuta y Melilla
Badajoz	Zaragoza
Ceuta y Melilla	Huelva
Salamanca	Tenerife
Jaén	Barcelona

5.3.3. Generalización

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de $\mathbf{S}$. Estas direcciones se denominan *direcciones principales* de los datos y a las nuevas variables por ellas definidas *componentes principales*. En general, la matriz $\mathbf{X}$ (y por tanto la $\mathbf{S}$) tiene rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características, $\lambda_1, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables, $\mathbf{S}$, mediante:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \tag{5.11}$$

y sus vectores asociados son:

$$(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{a}_i = 0. \tag{5.12}$$

Los términos λ_i son reales, al ser la matriz $\mathbf{S}$ simétrica, y positivos, ya que $\mathbf{S}$ es definida positiva. Por ser $\mathbf{S}$ simétrica si λ_j y λ_h son dos raíces distintas sus vectores asociados son ortogonales. Si $\mathbf{S}$ fuese semidefinida positiva de rango $r < p$, lo que ocurriría si $p - r$ variables fuesen combinación lineal de las demás, habría solamente r raíces características positivas y el resto serían ceros.

Llamando $\mathbf{Z}$ a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z} = \mathbf{XA}$$

donde $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Calcular los componentes principales equivale a aplicar una transformación ortogonal $\mathbf{A}$ a las variables $\mathbf{X}$ (ejes originales) para obtener unas nuevas variables $\mathbf{Z}$ incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los "ejes naturales" de los datos.

Ejemplo 5.4.
Los restantes valores propios de la matriz de covarianzas de los datos de la EPF son 0.027, 0.0175, 0.0126, 0.0107, 0.010, 0.0059, y 0.00526. A partir del tercero son muy pequeños. El tercer componente principal es

$$\begin{aligned} z_3 &= 0.12x_1 + 0.05x_2 + 0.34x_3 + 0.11x_4 - 0.85x_5 + 0.04x_6 - \\ &\quad - 0.30x_7 + 0.20x_8 + 0.003x_9 = \\ &= (0.12x_1 + 0.05x_2 + 0.34x_3 + 0.11x_4 + 0.04x_6 + 0.20x_8) - \\ &\quad - (0.85x_5 + 0.30x_7) \end{aligned}$$

y puede de nuevo interpretarse como la diferencia entre dos medias ponderadas. La primera da sobre todo peso a las variables 3, vivienda, 8, turismo y ocio, 1, alimentación y 4, mobiliario doméstico. La segunda a la 5, gastos sanitarios, y a la 7, enseñanza y cultura. Separa provincias con bajos costes en sanidad y altos en vivienda y ocio de las que tengan la

estructura opuesta. La Figura 5.3 representa las observaciones proyectadas sobre el plano de las componentes primera y tercera. Se observa que la tercera dimensión es independiente de la primera (riqueza o renta) y separa provincias con altos gastos en sanidad, como Salamanca y Palencia, de otras con gastos relativamente bajos en esta magnitud y mayor en vivienda y ocio.

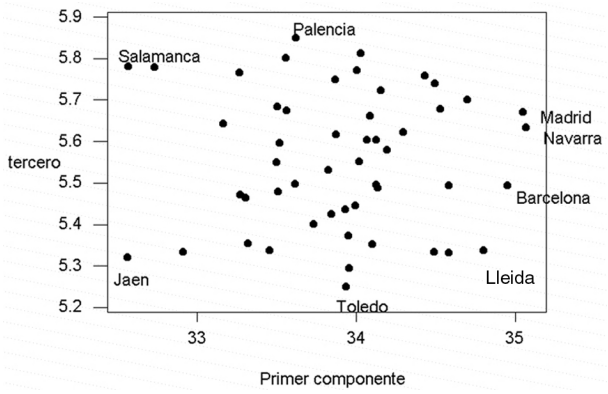


Figura 5.3. Representación de los datos de la EPF em el plano definido por los componentes primero y tercero.

Ejemplo 5.5.
La Tabla 5.2 presenta la matriz de varianzas y covarianzas entre nueve indicadores económicos medidos en distintas empresas.

Tabla 5.2. Matriz de varianzas covarianzas de los nueve indicadores

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
177	179	95	96	53	32	−7	−4	−3
	419	245	131	181	127	−2	1	4
		302	60	109	142	4	0,4	11
			158	102	42	4	3	2
				137	96	4	5	6
					128	2	2	8
						34	31	33
							39	39
								48

Los valores propios de esta matriz se presentan en la Tabla 5.3. Su suma es 1441.8, prácticamente igual, salvo por errores de redondeo, a la suma de las varianzas de las variables, que es 1442. Ya veremos que esta concordancia ocurre siempre. Los vectores propios de los

tres primeros componentes se indican en la Tabla 5.4. Se observa que el primer componente principal es una media ponderada de las primeras seis variables. El segundo contrapone la primera, la segunda y la cuarta a la tercera y la sexta. El tercer componente contrapone las tres primeras al resto de las variables.

Estos resultados son consistentes con la matriz de la Tabla 5.2. El rasgo más característico de esta tabla es la distinta magnitud de las seis primeras variables respecto al resto. Esto lo recoge el primer componente principal. El segundo rasgo es la presencia de covarianzas negativas en las filas de las dos primeras variables y esto se recoge en el segundo componente. El tercero incorpora por un lado las tres últimas variables y, por otro, contrapone las tres primeras variables frente al resto.

Tabla 5.3. Autovalores de la matriz Tabla 5.2

Componente	1	2	3	4	5	6	7	8	9
λ_i	878.5	196.1	128.6	103.4	81.2	37.8	7.0	5.7	3.5

Tabla 5.4. Vectores propios de la matriz Tabla 5.2

Componente	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.30	0.66	0.48	0.26	0.32	0.27	0.00	0.00	0.01
2	−0.48	−0.15	0.58	−0.49	−0.04	0.37	0.06	0.04	0.08
3	−0.41	−0.18	−0.23	0.45	0.49	0.27	0.26	0.28	0.29

5.4. Propiedades de los componentes

Los componentes principales son nuevas variables con las propiedades siguientes:

1. Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
   Comprobemos el primer punto. Como $Var\ (z_h) = \lambda_h$ y la suma de los valores propios es la traza de la matriz:

   $$tr(\mathbf{S}) = Var\ (x_1) + \ldots + Var\ (x_p) = \lambda_1 + \ldots + \lambda_p$$

   por tanto, $\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(z_i)$. Las nuevas variables z_i tienen conjuntamente la misma variabilidad que las variables originales.
   Los componentes principales también conservan la *Varianza generalizada*, (determinante de la matriz de covarianzas de las variables). Como el determinante

es el producto de los valores propios, llamando $\mathbf{S}_z$ a la matriz de covarianzas de los componentes, que es diagonal con términos λ_i :

$$|\mathbf{S}_x| = \lambda_1 \dots \lambda_p = \Pi_{i=1}^{p} Var(z_i) = |\mathbf{S}_z| .$$

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

   En efecto, la varianza del componente h es λ_h, y la suma de las varianzas de las variables originales es $\sum_{i=1}^{p} \lambda_i$, igual, como acabamos de ver, a la suma de las varianzas de los componentes. La proporción de variabilidad total explicada por el componente h es $\lambda_h / \sum \lambda_i$.

3. Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio:

   $$Cov(z_i; x_1, \dots, x_p) = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

   donde $\mathbf{a}_i$ es el vector de coeficientes de la componente z_i.
   Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

   $$Cov(z, x) = \frac{1}{n} \mathbf{Z}'\mathbf{X}$$

   y su primera fila proporciona las covarianzas entre la primera componente y las p variables originales. Como $\mathbf{Z} = \mathbf{XA}$, sustituyendo

   $$Cov(z, x) = \frac{1}{n} \mathbf{A}'\mathbf{X}'\mathbf{X} = \mathbf{A}'\mathbf{S} = \mathbf{DA}',$$

   donde $\mathbf{A}$ contiene en columnas los vectores propios de $\mathbf{S}$ y $\mathbf{D}$ es la matriz diagonal de los valores propios. En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $\mathbf{A}'\mathbf{S}$, es decir $\mathbf{a}_1'\mathbf{S}$ o también $\lambda_1 \mathbf{a}_1'$, donde $\mathbf{a}_1'$ es el vector de coeficientes de la primera componente principal.

4. Las correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.

   Para comprobarlo:

   $$Corr(z_i; x_j) = \frac{Cov(z_i x_j)}{\sqrt{Var(z_i) Var(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

5. Las r componentes principales (r < p) proporcionan la predicción lineal óptima con r variables del conjunto de variables X.

   Esta afirmación puede demostrarse de dos formas. La primera demostrando que la mejor predicción lineal con r variables de las variables originales se obtiene utilizando las r primeras componentes principales. La segunda demostrando que la mejor aproximación de la matriz de datos que puede construirse con una matriz de rango r se obtiene construyendo esta matriz con los valores de los r primeros componentes principales. La demostración de estas propiedades puede verse en el apéndice 5.1.

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

Estandarizando los componentes $\mathbf{Z}$ por sus desviaciónes típicas, se obtienen las nuevas variables

$$\mathbf{Y}_c = \mathbf{Z}\mathbf{D}^{-1/2} = \mathbf{X}\mathbf{A}\mathbf{D}^{-1/2}$$

donde $\mathbf{D}^{-1/2}$ es la matriz que contienen las inversas de las desviaciónes típicas de las componentes. Hemos visto en el capítulo anterior que la estandarización multivariante de una matriz de variables $\mathbf{X}$ de media cero se define como:

$$\mathbf{Y}_s = \mathbf{X}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{A}'$$

Tanto las variables $\mathbf{Y}_c$ como las $\mathbf{Y}_s$ tienen matriz de covarianzas identidad, pero unas pueden ser una rotación de las otras. Esto no altera sus propiedades, y la estandarización multivariante puede interpretarse como:

(1) obtener los componentes principales;

(2) estandarizarlos para que tengan todos la misma varianza.

Esta relación se presenta gráficamente en la Figura 5.4. La transformación mediante componentes principales conduce a variables incorreladas pero con distinta varianza. Puede interpretarse como rotar los ejes de la elipse que definen los puntos para que coincidan con sus ejes naturales. La estandarización multivariane produce variables incorreladas con varianza unidad, lo que supone buscar los ejes naturales y luego estandarizarlos. En consecuencia, si estandarizamos los componentes se obtiene las variables estandarizadas de forma multivariante.

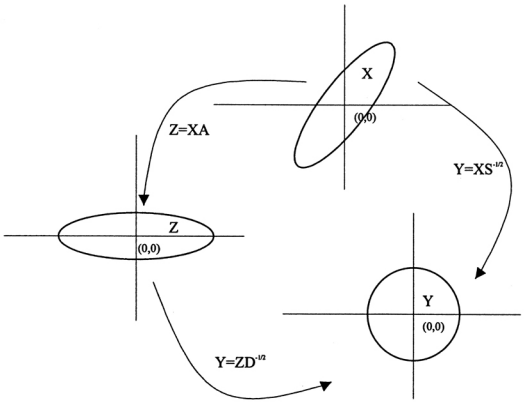


Figura 5.4. Representación gráfica de la relación entre componentes principales y estandarización multivariante.

5.5. Análisis normado o con correlaciones

Los componentes principales se obtienen maximizando la varianza de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j s_{ij} \tag{5.13}$$

con la restricción $\mathbf{a}'\mathbf{a} = 1$. Si alguna de las variables, por ejemplo la primera, tiene una varianza s_1^2, mayor que las demás, la manera de aumentar M es hacer tan grande como podamos la coordenada a_1 asociada a esta variable. En el límite, si una variable tiene una varianza mucho mayor que las demás, el primer componente principal coincidirá muy aproximadamente con esta variable.

Cuando las variables tienen unidades distintas esta propiedad no es conveniente: si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en magnitud sus valores numéricos (pasamos por ejemplo de medir en km a medir en metros), el peso de esa variable en el análisis aumentará, ya que en (5.13):

(1) su varianza será mayor y aumentará su coeficiente en el componente, a_i^2, pues contribuye más a aumentar M;

(2) sus covarianzas con todas las variables aumentarán, con el consiguiente efecto de incrementar a_i.

En resumen, cuando las escalas de medida de las variables son muy distintas, la maximización de (5.13) dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis. Si queremos evitar

este problema, conviene estandarizar las variables antes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables X sean similares.

La estandarización resuelve otro posible problema. Si las variabilidades de las X son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente. Este problema se evita al estandarizar las variables, ya que entonces las varianzas son la unidad, y las covarianzas son los coeficientes de correlación. La ecuación a maximizar se transforma en:

$$M' = 1 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j r_{ij} \tag{5.14}$$

siendo r_{ij} el coeficiente de correlación lineal entre las variables i y j. En consecuencia, la solución depende de las correlaciones y no de las varianzas.

Los *componentes principales normados* se obtienen calculando los vectores y valores propios de la matriz $\mathbf{R}$, de coeficientes de correlación. Llamando λ_p^R a las raíces características de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p \tag{5.15}$$

Las propiedades de los componentes extraídos de $\mathbf{R}$ son:

1. La proporción de variación explicada por λ_p^R será:

$$\frac{\lambda_p^R}{p} \tag{5.16}$$

2. Las correlaciones entre cada componente z_j y las variables X originales vienen dados directamente por $a'_j\sqrt{\lambda_j}$ siendo $\mathbf{z}_j = \mathbf{X}\mathbf{a}_j$.

Estas propiedades son consecuencia inmediata de los resultados de la Sección 5.4.

Cuando las variables X originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones o análisis normado. Cuando las variables tienen las mismas unidades, ambas alternativas son posibles. Si las diferencias entre las varianzas de las variables son informativas y queremos tenerlas en cuenta en el análisis, no debemos estandarizar las variables: por ejemplo, supongamos dos índices con la misma base pero uno fluctúa mucho y el otro es casi constante. Este hecho es informativo, y para tenerlo en cuenta no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso. Por el contrario, si las diferencias de variabilidad no son relevantes se eliminan con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquél que conduzca a conclusiones más informativas.

Ejemplo 5.6.
La matriz de correlación de los nueve indicadores económicos del Ejemplo 5.5 es

$$\mathbf{R} = \begin{bmatrix} 1 & 0.66 & 0.41 & 0.57 & 0.34 & 0.21 & -0.09 & -0.05 & -0.03 \\ & 1 & 0.69 & 0.51 & 0.76 & 0.55 & -0.01 & 0.01 & 0.03 \\ & & 1 & 0.28 & 0.54 & 0.72 & 0.04 & 0.00 & 0.09 \\ & & & 1 & 0.69 & 0.30 & 0.05 & 0.03 & 0.02 \\ & & & & 1 & 0.73 & 0.06 & 0.07 & 0.07 \\ & & & & & 1 & 0.03 & 0.03 & 0.10 \\ & & & & & & 1 & 0.85 & 0.82 \\ & & & & & & & 1 & 0.90 \\ & & & & & & & & 1 \end{bmatrix}$$

Los valores propios son:

λ_i	3.70	2.72	1.06	0.70	0.30	0.23	0.16	0.09	0.03

y los vectores propios asociados a los tres primeros valores propios son:

Tabla 5.5. Vectores propios de la matriz de correlaciones

λ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
3.7	0.34	0.46	0.41	0.36	0.46	0.40	0.06	0.06	0.08
2.72	−0.11	−0.07	−0.03	−0.04	−0.02	−0.01	0.56	0.58	0.57
1.06	−0.54	−0.05	0.38	−0.52	0.07	0.53	−0.04	−0.07	0.00

Si comparamos estos resultados con los del Ejemplo 5.5 vemos que el primer vector propio cambia apreciablemente. Con la matriz de varianzas las variables con más peso en el componente eran las que tenían una mayor varianza: la 2, luego la 3 y finalmente las 1, 4, 5 y 6 con un peso parecido. Estos pesos reproducen la relación relativa entre las varianzas de las variables. Sin embargo, al utilizar la matriz de correlaciones este efecto desaparece, y el peso de las variables está más relacionado con las correlaciones. La proporción de variabilidad explicada por el primer componente cambia mucho: de $878.5/1441.8 = 60.9$ por 100 a $3.7/9 = = 41$ por 100.

El segundo componente cambia completamente: ahora está prácticamente asociado a las tres últimas variables. La proporción de variabilidad que explica ha aumentado considerablemente, del $196/1441.8 = 13.6$ por 100 a $2.72/9 = 30$ por 100. El tercer vector propio es también distinto en ambas matrices.

Ejemplo 5.7.
Consideremos los datos de INVEST publicaciones científicas en los países de la OCDE. Los datos tienen magnitudes muy distintas (unos bancos de datos tienen muchos más trabajos que otros). Si deseamos conservar esta propiedad, que está asociada a que en algunos campos científicos se publica mucho más que en otros, haremos el análisis sobre la matriz de covarianzas. Si no queremos dar más peso a unos campos que a otros, es conveniente realizar

el análisis normado o sobre la matriz de correlación. Los resultados en este último caso se indican en la Tabla 5.6.

Se observa que el primer componente principal explica una proporción muy alta de la variabilidad, el 95.4 por 100. Con los tres primeros componentes se explica el 99.5 por 100 de la variabilidad. Además, después del tercer vector propio la variabilidad explicada disminuye claramente, (véase la Tabla 5.6 y la Figura 5.5 lo que indica que sólo debemos preocuparnos de los tres primeros componentes ya que los siguientes tienen poca capacidad explicativa. En la Tabla 5.7 se indican los valores de los componentes para estos tres vectores propios.

Tabla 5.6. Variabilidad explicada por los componentes principales para los datos de INVEST en logaritmos

Comp.	λ_h	P_h	$\sum_{i=1}^{h} P_h$
1	7.630	0.954	0.954
2	0.207	0.026	0.980
3	0.121	0.015	0.995
4	0.019	0.002	0.997
5	0.017	0.002	0.999
6	0.004	0.001	1.000
7	0.001	0.000	1.000
8	0.000	0.000	1.000

Tabla 5.7. Vectores propios de los tres primeros componentes para los datos de INVEST en logaritmos

	Comp. 1	Comp. 2	Comp. 3
INTER.A	0.358	−0.173	0.36
INTER.F	0.360	−0.098	0.08
AGRIC.	0.355	−0.366	−0.10
BIOLO.	0.346	−0.359	−0.69
MEDIC.	0.361	−0.070	0.15
QUÍMI.	0.334	0.786	−0.41
INGEN.	0.354	0.268	0.40
FÍSICA	0.361	0.054	0.17

Para interpretar los componentes consideramos sus coordenadas en las variables. Éstas se indican en la Tabla 5.7 y en la Figura 5.6. Se observa que el primer componente es un factor de tamaño, ya que es una media ponderada de todas las variables con mayor peso de los bancos interdisciplinarios y del banco médico. El segundo componente es un factor de forma y contrapone la investigación en Química e Ingeniería frente a la realizada en Agricultura y Biología. El tercero contrapone ingeniería, física y el banco interA con respecto a Biología y Química.

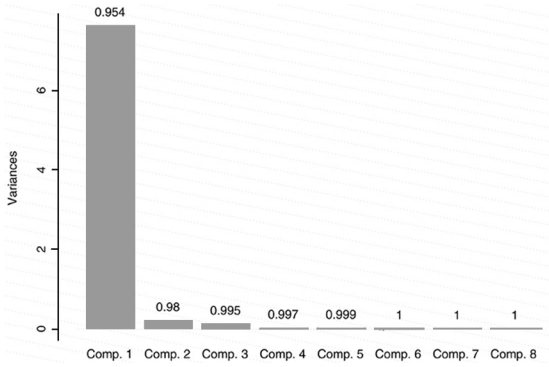


Figura 5.5. Gráfico para la selección del número de componentes. Datos de INVEST en logaritmos.

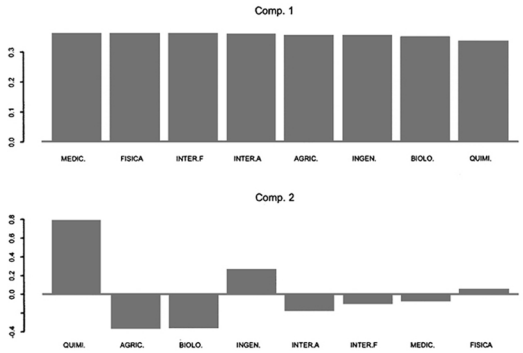


Figura 5.6. Representación de los pesos de las dos componentes. Datos de INVEST en logaritmos.

5.6. Interpretación de los componentes

Componentes de tamaño y forma

Cuando existe una alta correlación positiva entre todas las variables, el primer componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables (véase el Ejercicio 5.2), o un factor global de "tamaño". Los restantes componentes se interpretan como factores "de forma" y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de variables frente a otros. Estos factores de forma pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las variables de un signo a las del otro. Por ejemplo,

el segundo componente principal de los datos de la EPF del Ejercicio 5.3 puede escribirse aproximadamente, despreciando los coeficientes pequeños (menores que 0,1):

$$z_2 = (0.05x_1 + 0.16x_2 + 0.07x_4 + 0.29x_6 + 0.78x_9) - \\ - (0.17x_3 + 0.21x_5 + 0.40x_7 + 0.17x_8) \simeq I_0 - I_S$$

donde

$$I_0 = 0.16x_2 + 0.29x_6 + 0.78x_9$$

es un indicador de los gastos de transporte y transferencias a otras provincias y

$$I_S = 0.17x_3 + 0.21x_5 + 0.40x_7 + 0.17x_8$$

es un indicador de gastos en servicios (educación y sanidad). Además, cuando las variables van en logaritmos, los componentes suelen poder escribirse como ratios de promedios geométricos de las variables. Por ejemplo, supongamos que un componente tiene la expresión

$$z_1 = -0.5 \log x_1 + 0.3 \log x_2 + 0.2 \log x_3$$

este componente puede escribirse también como

$$z_1 = 0.3 \log \frac{x_2}{x_1} + 0.2 \log \frac{x_3}{x_1}$$

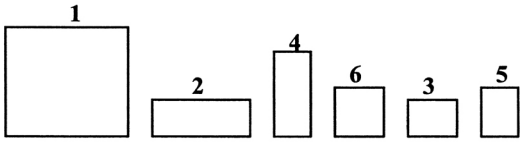
que indica que es un promedio de estos dos ratios (véase el Ejemplo 5.1).

La interpretación de los componentes se simplifica suponiendo que los coeficientes pequeños son cero y redondeando los coeficientes grandes para expresar el componente como cocientes, diferencias o sumas entre variables. Estas aproximaciones son razonables si modifican poco la estructura del componente y mejoran su interpretación. Una medida del cambio introducido al modificar un vector propio de $\mathbf{a}_i$ a $\mathbf{a}_{iM}$ es el cambio en la proporción de variabilidad explicada por el componente. Si el valor propio asociado a $\mathbf{a}_i$ es λ_i, el componente explica el $\lambda_i / \sum \lambda_j$ de la variabilidad. Si ahora modificamos el vector a $\mathbf{a}_{iM}$, la varianza de la proyección de los datos sobre este componente es $\lambda_{iM} = \mathbf{a}'_{iM}\mathbf{S}\mathbf{a}_{iM} = (\widetilde{\mathbf{X}}\mathbf{a}_{iM})'(\widetilde{\mathbf{X}}\mathbf{a}_{iM})/n$, la varianza del componente, y la proporción de variabilidad explicada será $\lambda_{iM} / \sum \lambda_j$. El cambio relativo será $(\lambda_i - \lambda_{iM})/\lambda_i$, ya que siempre $\lambda_i \geq \lambda_{iM}$, y si este cambio es pequeño, está justificada la modificación si favorece la interpretación.

Ejemplo 5.8.
Vamos a calcular el cambio relativo que experimenta el segundo componente principal de los datos de la EPF si despreciamos los coeficientes más pequeños. La varianza del segundo componente modificado es 0.0319. La varianza del componente original es 0.0320, por lo que el cambio de explicación por tomar el coeficiente simplificado es sólo de

$$(0.0320 - 0.0319)/0.0320 = 1/320 = 0.0031.$$

Ejemplo 5.9.
Supongamos 6 observaciones $x_1, \ldots, x_6$ en dos dimensiones, cada observación corresponde a un rectángulo y las variables son longitud de la base y altura del rectángulo. Gráficamente las observaciones son,



que corresponden a la matriz de datos,

$$\mathbf{X} = \begin{bmatrix} 2 & 2 \\ 1.5 & 0.5 \\ 0.7 & 0.5 \\ 0.5 & 1.5 \\ 0.5 & 0.7 \\ 0.7 & 0.7 \end{bmatrix}$$

aplicamos logaritmos a estos datos para facilitar la interpretación de las componentes,

$$\log(\mathbf{X}) = \begin{bmatrix} 0.301 & 0.301 \\ 0.176 & -0.301 \\ -0.155 & -0.301 \\ -0.301 & 0.176 \\ -0.301 & -0.155 \\ -0.155 & -0.155 \end{bmatrix}$$

cuya matriz de varianzas covarianzas es,

$$\mathbf{S} = \begin{bmatrix} 6.39 & 1.41 \\ 1.41 & 6.39 \end{bmatrix} .10^{-2}$$

Los autovalores y autovectores de la descomposición espectral de esta matriz son,

$$\lambda_1 = 0.78 \qquad \lambda_2 = 0.0498$$

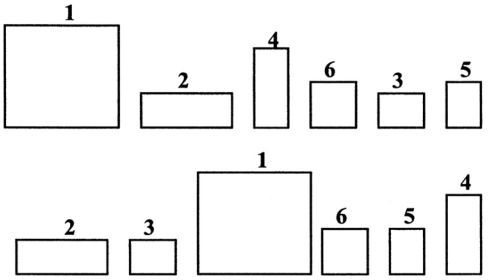
$$a_1 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \qquad a_2 = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix}$$

las dos primeras componentes son

$$Z_1 = Xa_1 = 0.707\log(X_1) + 0.707\log(X_2) = 0.707\log(X_1X_2) = \begin{bmatrix} 0.426 \\ -0.088 \\ -0.322 \\ -0.088 \\ -0.322 \\ -0.219 \end{bmatrix}$$

$$Z_2 = Xa_2 = 0.707\log(X_1) - 0.707\log(X_2) = 0.707\log(\frac{X_1}{X_2}) = \begin{bmatrix} 0 \\ 0.337 \\ 0.103 \\ -0.337 \\ -0.103 \\ 0 \end{bmatrix}$$

Si ordenamos los rectángulos según el valor de la primera y segunda componente obtenemos,



La primera ordenación coincide con la inducida por el volumen de los rectángulos, es una transformación creciente del producto de la base por la altura, y el primer componente describe el tamaño. El segundo componente relaciona la base con la altura y ordena las observaciones en función de su forma.

5.6.1. Selección del número de componentes

Se han sugerido distintas reglas para seleccionar el número de componentes:

1. Realizar un gráfico de λ_i frente a i. Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i. La idea es buscar un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

2. Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80 o el 90 por 100. Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo, es posible que un único componente de "tamaño" recoja el 90 por 100 de la variabilidad y, sin embargo, pueden existir otros componentes que sean muy adecuados para explicar la "forma" de las variables.

3. Desechar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media, $\sum \lambda_i/p$. En particular, cuando se trabaja con la matriz de correlación, el valor medio de los componentes es 1, y esta regla lleva a seleccionar los valores propios mayores que la unidad. De nuevo esta regla es arbitraria: una variable que sea independiente del resto suele llevarse un componente principal (véase Ejercicio 5.8) y puede tener un valor propio mayor que la unidad. Sin embargo, si está incorrelada con el resto puede ser una variable poco relevante para el análisis, y no aportar mucho a la comprensión del fenómeno global.

5.6.2. Representación gráfica

La interpretación de los componentes principales se favorece representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de los componentes principales más importantes. Este punto se ha ilustrado en los ejemplos anteriores, donde se ha indicado que la proyección de cualquier observación sobre un componente es directamente el valor del componente para esa observación. La representación habitual es tomar dos ejes ortogonales que representen los dos componentes considerados, y situar cada punto sobre ese plano por sus coordenadas con relación a estos ejes, que son los valores de los dos componentes para esa observación. Por ejemplo, en el plano de los dos primeros componentes, las coordenadas del punto $\mathbf{x}_i$ son $z_{1i} = \mathbf{a}_1'\mathbf{x}_i$ y $z_{2i} = \mathbf{a}_2'\mathbf{x}_i$.

La interpretación se favorece representando en el mismo plano, además de las observaciones, las variables originales. Esto puede hacerse utilizando como coordenadas su coeficiente de correlación con cada uno de los ejes. El vector de correlaciones entre el primer componente y las variables originales viene dado por $\lambda_1^{1/2}\mathbf{a}_1'\mathbf{D}$, donde $\mathbf{D}$ es una matriz diagonal cuyos términos son las inversas de las desviaciones típicas de cada variable. La matriz de correlaciones $\mathbf{R}_{cv}$ entre los p componentes y las p variables tendrá como filas los términos $\lambda_j^{1/2}\mathbf{a}_j'\mathbf{D}$ y puede escribirse

$$\mathbf{R}_{cv} = \mathbf{\Lambda}^{1/2}\mathbf{A}\mathbf{D}$$

donde $\mathbf{A}$ es la matriz de vectores propios, $\mathbf{\Lambda}^{1/2}$ es la matriz diagonal con términos $\sqrt{\lambda_i}$ y en el análisis normado como las variables se estandarizan a varianza unidad las correlaciones será simplemente $\mathbf{\Lambda}^{1/2}\mathbf{A}$.

Una representación equivalente es el biplot que presentamos en la sección siguiente. Tiene la ventaja de representar las variables y las observaciones en un mismo gráfico.

Conviene investigar si transformando las variables se obtiene una interpretación más simple. Como regla general, cuando al tomar logaritmos las variables $\mathbf{X}$ tienen una distribución aproximadamente simétrica, conviene realizar el análisis de componentes principales sobre los logaritmos de las variables.

Es importante recordar que las covarianzas (o correlaciones) miden únicamente las relaciones lineales entre las variables. Cuando entre ellas existan relaciones fuertes no lineales el análisis de componentes principales puede dar una información muy parcial de las variables.

Ejemplo 5.10.
La Figura 5.7 presenta la proyección de los datos de INVEST, los países de la OCDE, sobre el plano formado por los dos primeros componentes principales extraídos de la matriz de correlación, que se estudiaron en el Ejemplo 5.6. Se observa que el primer eje ordena a los países por su cantidad de investigación, mientras que el segundo tiene en cuenta sus características: separa a Japón (JP), con gran énfasis en investigación tecnológica, del Reino Unido (UK), que tiene más énfasis en la investigación biomédica.

Como indicamos en el Capítulo 4 la observación de Estados Unidos es atípica y existe una marcada asimetría en las distribuciones de las variables. Vamos a presentar los datos excluyendo a Estados Unidos y con una transformación logarítmica de las variables para reducir la asimetría. La Figura 5.8 muestra el nuevo diagrama de cajas múltiple. Como la varianza de las nuevas variables transformadas es similar, el análisis de componentes principales se realizará directamente sobre la matriz de varianzas covarianzas. Los resultados obtenidos figuran en las Tablas 5.8 y 5.9.

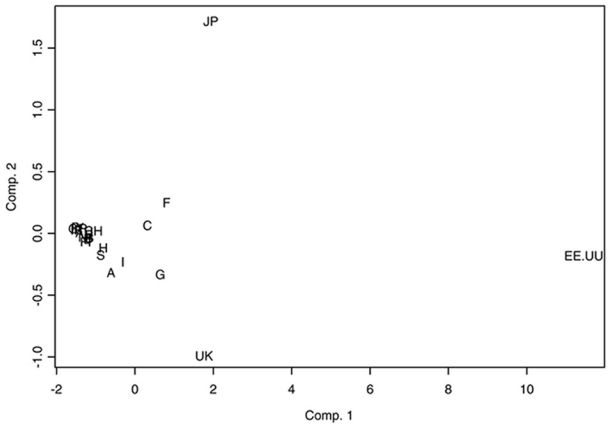


Figura 5.7. Proyección de las observaciones en las dos primeras componentes principales.

Los tres primeros componentes explican el 97 por 100 de la variabilidad y tienen la siguiente interpretación. El primero, es una media ponderada de todos los bancos con mayor peso del banco químico. El segundo, contrapone la investigación en Química frente a la general del banco INTER.F y a la de ingeniería y física. El tercero contrapone el banco INTER.F y Química al resto.

Los países proyectados en estos tres componentes se presentan en la Figura 5.9. Se ha añadido también la proyección sobre el cuarto componente, que separa completamente al Reino Unido de Japón (véase Tabla 5.9).

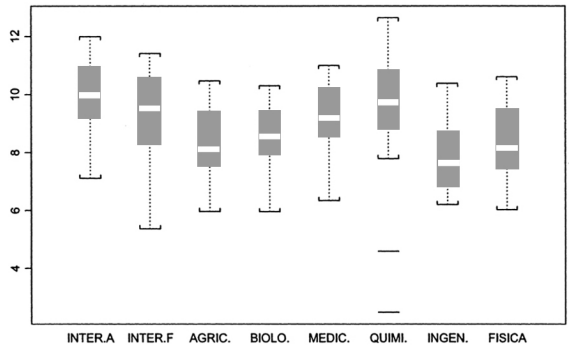


Figura 5.8. Diagrama de cajas de los logaritmos de las variables de INVEST una vez eliminado Estados Unidos.

Tabla 5.8. Variabilidad explicada por los componentes principales

	λ_h	P_h	$\sum_{i=1}^{h} P_h$
Comp. 1	14.98	0.90	0.90
Comp. 2	0.83	0.05	0.94
Comp. 3	0.50	0.03	0.97
Comp. 4	0.21	0.01	0.99
Comp. 5	0.10	0.01	0.99
Comp. 6	0.08	0.00	1.00
Comp. 7	0.02	0.00	1.00
Comp. 8	0.02	0.00	1.00

Tabla 5.9. Pesos de las tres primeras componentes principales para los datos de INVEST

	Comp. 1	Comp. 2	Comp. 3
INTER.A	0.31	0.05	−0.40
INTER.F	0.37	0.63	0.63
AGRIC.	0.30	0.07	−0.14
BIOLO.	0.27	−0.06	−0.30
MEDIC.	0.32	0.01	−0.25
QUÍMI.	0.56	−0.70	0.41
INGEN.	0.28	0.25	−0.18
FÍSICA	0.32	0.21	−0.26

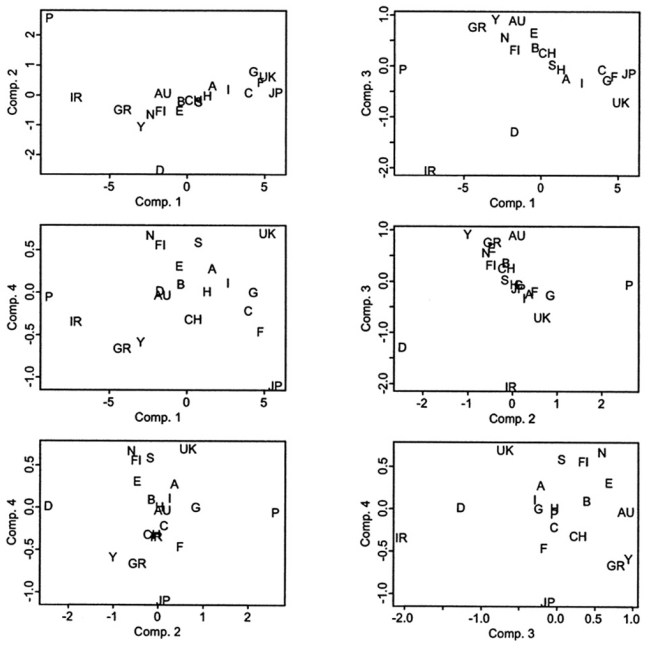


Figura 5.9. Representación de las observaciones de INVEST en los planos definidos por las cuatro primeras componentes.

5.6.3. Datos atípicos

Antes de obtener los componentes principales conviene asegurarse de que no existen datos atípicos, ya que, como hemos visto en el capítulo anterior, los atípicos pueden distorsionar totalmente la matriz de covarianzas.

Para ilustrar su efecto sobre los componentes, supongamos el caso más simple en que un error de medida introduce un valor atípico grande en la primera variable. Su efecto será aumentar mucho la varianza de esta variable y disminuir las covarianzas con las restantes, con lo que, si hacemos el atípico muy grande, la matriz $\mathbf{S}$ será, aproximadamente:

$$\begin{bmatrix} \sigma_1^2 & \mathbf{0}' \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}$$

donde $\mathbf{0}'=(0,0,...,0)$. Esta matriz tiene un vector propio $(1, 0, \ldots, 0)$ unido al valor propio σ_1^2 y, si σ_1^2 es muy grande, éste será el primer componente principal. Por tanto, un valor atípico suficientemente grande distorsiona todos los componentes que podemos obtener de la matriz afectada (véase el Ejemplo 5.10).

El resultado anterior sugiere que las componentes principales podrían utilizarse para detectar datos atípicos multivariantes, ya que un valor muy extremo se llevará un componente principal y aparecerá como extremo sobre esta componente. Desgraciadamente, aunque los componentes pueden identificar atípicos aislados, no hay garantía de que funcionen cuando existen grupos de atípicos, debido al problema de enmascaramiento. Por esta razón, conviene utilizar para detectarlos el método presentado en el capítulo anterior, basado en proyecciones sobre las direcciones extremas de kurtosis, que al ser capaz de identificar todos los posibles atípicos permite calcular la matriz de covarianzas libre de distorsiones graves.

5.6.4. Distribución de los componentes

Los componentes principales pueden verse como un conjunto nuevo de variables y estudiar su distribución individual y conjunta. Por construcción estarán incorrelados, pero pueden existir relaciones no lineales entre ellos.

Ejemplo 5.11.
Vamos a calcular los componentes principales de la matriz de correlación de las 27 medidas físicas, MEDIFIS. Aunque todas las variables van en centímetros, los tamaños de las variables son muy distintos, lo que aconseja utilizar la matriz de correlación. La proporción de varianza que explica cada vector propio se indica en la Tabla 5.10.

Para decidir cuántos componentes tomar utilizaremos la Figura 5.10 que indica que a partir del tercer componente hay una caída en la capacidad predictiva. Los tres primeros componentes explican conjuntamente el 93.5 por 100 de la variabilidad.

Tabla 5.10. Variabilidad explicada por las componentes para los datos de Medifis

λ_h	5.56	0.62	0.39	0.17	0.14	0.10	0.05
$P_h\%$	78.96	8.87	5.65	2.48	1.98	1.37	0.68

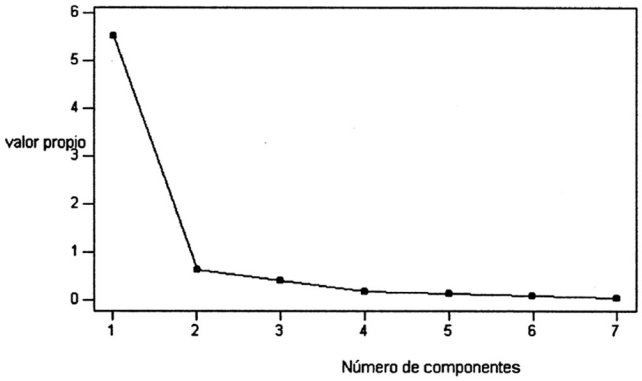


Figura 5.10. Gráfico para seleccionar el número de componentes, datos de MEDIFIS.

Los tres primeros vectores propios son:

	est	pes	pie	lbr	aes	dcr	drt
Comp. 1	.41	.39	.40	.39	.38	.29	.37
Comp. 2	−.16	.04	−.20	−.30	.11	.89	−.15
Comp. 3	.04	−.29	.13	−.15	−.57	.20	.71

El primer componente es una media de todas las medidas físicas y, por tanto, una medida del tamaño del cuerpo, siendo la variable con menor peso el diámetro del cráneo. El segundo es de forma, y está dominado por el diámetro del cráneo, variable poco correlada con el resto y, que determina por sí sola un componente principal, al no poder explicarse como combinación de otras. El tercer componente diferencia longitud frente a anchura: da mayor peso a la longitud de la pierna (drt) y lo contrapone al peso y a la anchura de la espalda.

La Figura 5.11 presenta un gráfico de las observaciones sobre el plano de los dos primeros componentes principales. Las coordenadas son las puntuaciones estandarizadas $z_i^* = \mathbf{X}^*\mathbf{a}_i$, $i = 1, 2$, donde $\mathbf{X}^*$ es la matriz de variables estandarizadas (de media cero y varianza uno). En este gráfico cada punto se indica con un 1, cuando la observación corresponde a un hombre y un 0 cuando es mujer. Puede verse que la primera componente de "tamaño" separa casi perfectamente los hombres de las mujeres. El segundo componente no parece reflejar ningún efecto del sexo. La primera componente es capaz, por sí misma, de explicar casi el 80 por 100 de variabilidad. Dado que el diámetro del cráneo está poco correlado con el resto de las variables, siendo casi en exclusiva responsable de la segunda dimensión, vamos a repetir el análisis eliminando esta variable.

Los resultados al eliminar la variable diámetro del cráneo se presentan en la tabla siguiente, que incluye los dos primeros valores y vectores propios que explican el 92 por 100 de la variabilidad.

λ_h	$P_h\%$	est	pes	pie	lbr	aes	drt
5.1	85	.43	.41	.42	.41	.39	.38
.4	7	.08	−.32	.17	−.04	−.60	.71
$Corr(z_1x_i)$		.97	.93	.95	.93	.88	.86
$Corr(z_2x_i)$		.05	−.20	.11	−.030	−.38	.45

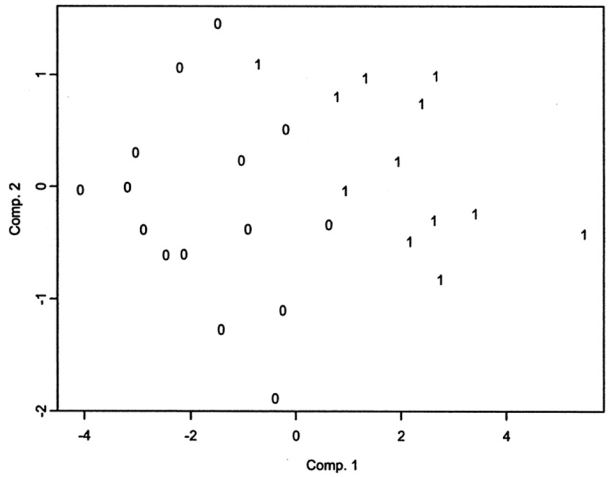


Figura 5.11. Proyección de las observaciones en las dos primeras componentes principales. Datos de MEDIFIS.

El primer componente es de nuevo una media ponderada que indica el tamaño de las personas, dando el mayor peso a la estatura de la persona. El segundo es de forma, ya que contrapone la longitud de la pierna a la anchura de la espalda y tiene peso positivo en las longitudes (del pie y estatura), y negativo en el peso. La proyección de los datos sobre el plano definido por los dos componentes se presenta en la Figura 5.12. Se observa que el primer componente de "tamaño" separa como antes los hombres de las mujeres, y que el segundo componente, al ser ortogonal al tamaño, no parece depender del sexo. Este componente separa, para ambos sexos, personas con constitución delgada de gruesa.

La Figura 5.13 presenta de forma gráfica las correlaciones entre el primer y segundo componente y cada variable, calculadas como $\sqrt{\lambda_h a_{hj}}$. Se observa que el primer componente está correlado con la altura y las restantes longitudes, mientras que el segundo está especialmente relacionado con la longitud de la pierna y la anchura de la espalda.

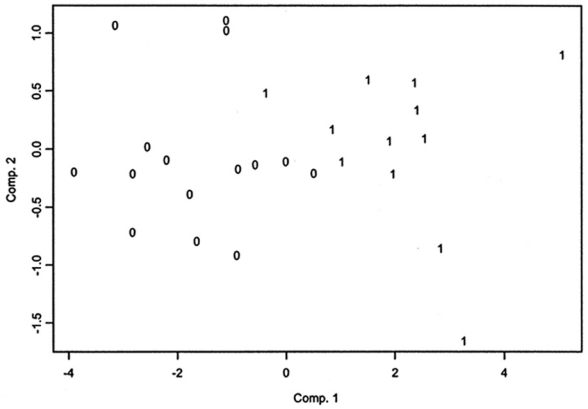


Figura 5.12. Proyección de las observaciones en las dos primeras componentes principales. Datos de MEDIFIS sin la variable diámetro del cráneo.

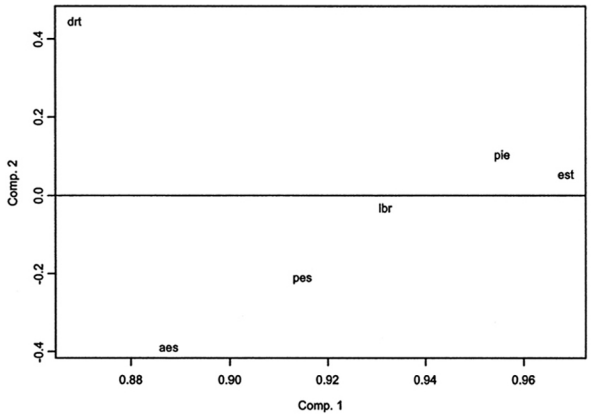


Figura 5.13. Correlación de las variables con las componentes principales. Datos de MEDIFIS.

Ejemplo 5.12.
Vamos a analizar la base de datos de MUNDODES (fichero mundodes.dat y Anexo I). Esta matriz de datos está constituida por 91 países en los que se han observado 9 variables: X_1: ratio de natalidad, X_2: ratio de mortalidad, X_3: mortalidad infantil, X_4: esperanza de vida en hombres, X_5: esperanza de vida de mujeres y X_6: PNB *per capita*.

La representación gráfica de las variables dos a dos, presentada en el capítulo anterior, muestra relaciones claramente no lineales. Aplicando transformaciones logarítmicas a las variables mejoramos la linealidad en estas relaciones.

Como las variables están medidas en distintas unidades se debe realizar un análisis de componentes principales normado (basado en la matriz de correlaciones), los resultados se presentan en la Figura 5.14.

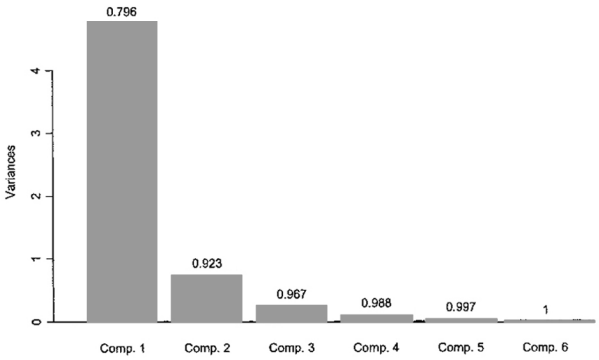


Figura 5.14. Proporción de variabilidad explicada por cada componente para los datos de MUNDODES.

La Figura 5.15 presenta el gráfico en forma de codo para seleccionar el número de componentes. El primer valor propio es 4.7278, y explica el 78.8 por 100 de la variabilidad. El segundo es 0.7261, y explica el 12 por 100. Hay un valor propio de 0,002 que corresponde a una variable que es prácticamente constante. Los vectores propios se presentan a continuación.

	PC1	PC2	PC3	PC4	PC5	PC6
X_1	−0.454	0.034	−0.130	0.159	0.378	0.780
X_2	0.416	0.196	0.513	0.683	0.233	0.067
X_3	0.341	−0.680	−0.524	0.307	0.225	−0.031
X_4	0.440	−0.052	0.222	−0.632	0.578	0.145
X_5	−0.452	0.085	−0.029	0.114	0.639	−0.605
X_6	−0.326	−0.699	0.628	−0.039	−0.100	0.002

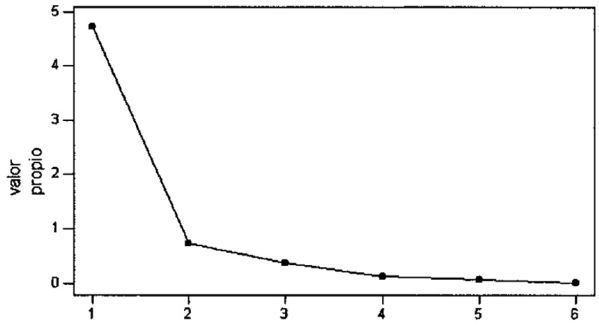


Figura 5.15. Gráfico de codo para el número de componentes.

El primer componente explica el 79% de la variabilidad, el segundo corresponde a un valor propio inferior a 1, pero lo incluiremos para interpretarlo. El primero se puede interpretar como una medida de desarrollo de un país, dado que las variables con peso positivo son las esperanzas de vida de hombres y mujeres y la renta, mientras que las de peso negativo son la mortalidad infantil y las tasas de natalidad y mortalidad, que son bajas en los países más desarrollados. El segundo componente está asociado a la mortalidad infantil y a la renta, con lo que resulta de difícil interpretación ya que mide una dimensión que está incorrelada con el primer término de desarrollo. Para interpretarla, la Figura 5.16 muestra los países en el plano de los dos componentes. Se observa que existe una fuerte relación no lineal entre los componentes, que aunque incorrelados no son claramente independientes. El primero ordena a los países por desarrollo y el segundo tiene en cuenta la mortalidad infantil y presenta una relación no lineal con la renta.

Los diagramas de dispersión mostraron relaciones entre las variables no lineales, por lo que vamos a repetir el análisis con las variables en logaritmos. Los valores propios de la matriz de correlaciones de las variables en logaritmos no cambian mucho, pero los vectores propios sí lo hacen. Son ahora:

	PC1	PC2	PC3	PC4	PC5	PC6
$\log X_1$	0.403	0.435	−0.376	−0.436	−0.562	0.033
$\log X_2$	0.307	−0.831	0.011	−0.457	−0.077	−0.020
$\log X_3$	0.433	0.267	−0.023	−0.331	0.793	0.051
$\log X_4$	−0.441	0.147	0.224	−0.531	0.019	−0.672
$\log X_5$	−0.446	0.071	0.213	−0.454	−0.012	0.738
$\log X_6$	−0.403	−0.149	−0.873	−0.057	0.223	−0.008

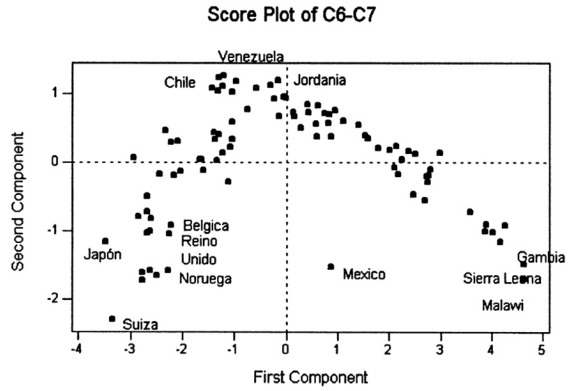


Figura 5.16. Representacion de los dos primeros componentes para los datos de MUNDODES.

El primero sigue siendo una medida de desarrollo pero ahora el segundo esta sobre todo ligado a la tasa de mortalidad. Separa países con alta y baja tasa de mortalidad. El último vector propio también tiene una interesante interpretación: indica que la diferencia en logaritmos entre las esperanzas de vida de hombres y mujeres es prácticamente constante en todos los países, ya que el valor propio que corresponde a este vector propio es muy pequeño (0,015). Los pesos asociados a cada una de las variables se presentan en la Figura 5.17.

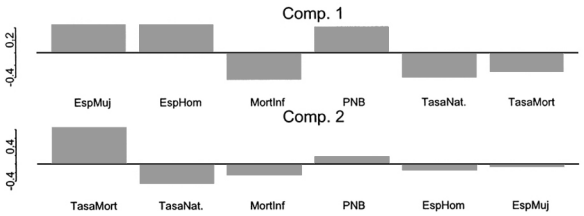


Figura 5.17. Pesos de las variables en los dos primeros componentes para los datos de MUNDODES.

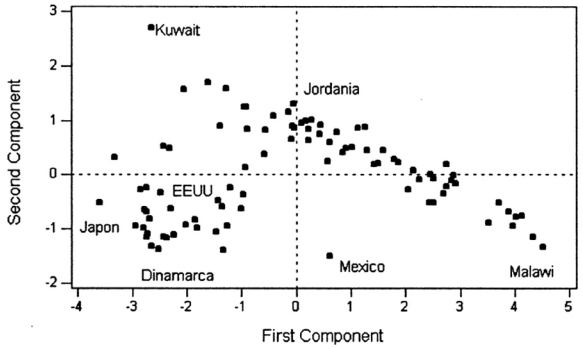


Figura 5.18. Gráfico de los datos de MUNDODES sobre los dos primeros componentes principales de los datos en logaritmos.

La Figura 5.18 representa los países en los dos primeros componentes. El primero, es una medida del desarrollo y el segundo depende principalmente de la tasa de mortalidad, y separa países que tienen alto (o bajo) valor aparente de desarrollo de otros que tienen una mortalidad mucho mayor de la que correspondería de acuerdo a su nivel de desarrollo. Ambas dimensiones están incorreladas, pero no son independientes, como se observa en la figura. Sin embargo, el grado de dependencia entre las variables es menor que con las variables sin transformar.

5.7. Generalizaciones

La idea de componentes principales puede extenderse para buscar representaciones no lineales de los datos que expliquen su estructura. Este enfoque es especialmente interesante si sospechamos que los datos pueden disponerse siguiendo una determinada superficie en el espacio. Como hemos visto, los vectores propios ligados a valores propios próximos a cero son muy importantes, porque revelan relaciones de poca variabilidad de los datos. Por ejemplo, supongamos para simplificar una variable bidimensional donde, aproximadamente, $f(x_1) + f(x_2) = c$. Entonces, si hacemos componentes principales de las cuatro variables $(x_1, x_2, f(x_1), f(x_2),)$ encontraremos un valor propio muy próximo a cero con un vector propio de la forma (0,0, 1,1).

Generalizando esta idea, si existe una relación cualquiera no lineal entre las variables, como esta relación podemos aproximarla por una relación polinómica

$$f(x_1, ..., x_p) = \sum a_i x_i + \sum b_{ij} x_i x_j + \sum c_{ijk} x_i x_j x_k + ...$$

si incluimos nuevas variables adicionales como $\mathrm{x}_1^2, ..., x_p^2$ o productos de variables $x_1 x_2$, etc., y extraemos los componentes principales de la matriz de correlaciones entre todas estas variables, si los puntos tienen una relación no lineal ésta se detectará ligada a un valor propio próximo a cero. Este enfoque se conoce a veces como *componentes principales generalizados*, y el lector interesado puede encontrar ejemplos de su aplicación en Gnandesikan (1977). El inconveniente de introducir nuevas variables, transformaciones de las iniciales, es que inmediatamente aumenta mucho la dimensión del problema, con lo que si la muestra no es muy grande podemos tener una matriz de correlaciones singular. Por otro lado, la interpretación de los resultados de este análisis, salvo en casos muy especiales, no suele ser fácil, con lo que esta herramienta se utiliza poco en la exploración de datos multivariantes.

5.8. Lecturas complementarias

Todos los textos generales de análisis multivariante que se indican en las referencias al final del libro incluyen componentes principales. Johnson y Wichern (1998) y Rechner (1998) son buenas presentaciones, con similar filosofía a la utilizada en este libro, mientras que Flury (1997) presenta un enfoque complementario al aquí expuesto. Componentes principales es un caso particular de los métodos de proyección introducidos en la Sección 4.2.3 que se conocen como *Projection Pursuit* (Búsqueda de la Proyección). Véase Krzanowski y Marriot (1994) para más detalles. Un excelente tratado sobre componentes principales y sus extensiones es el libro de Jackson (1991), que contiene numerosas referencias. La idea de componentes principales puede extenderse al caso no lineal, véase Gnanadesikan (1997), Salinelli (1998) y Delicado (2001). Los componentes principales puede aplicarse para investigar si varios grupos de datos tienen componentes comunes. Este aspecto ha sido investigado por Krzanowski (1988) y Flury (1988). Cuadras, C. M. (1991) y Aluja, T. y Morineau, A. (1999) son buenas referencias en español.

Ejercicios

5.1. Dada la matriz de covarianzas

$$\mathbf{S} = \begin{bmatrix} 1+d & 1 & 1 \\ 1 & 1+d & 1 \\ 1 & 1 & 1+d \end{bmatrix}$$

encontrar los componentes principales. Calcular la proporción de variabilidad explicada por cada uno y las correlaciones entre los componentes y las variables. Interpretar los componentes en función del tamaño de d.

5.2. Dada la matriz de correlación:

$$\mathbf{S} = \begin{bmatrix} 1 & d & d & d \\ d & 1 & d & d \\ d & d & 1 & d \\ d & d & d & 1 \end{bmatrix}$$

encontrar la primera componente principal. (Nota, utilizar que $\sum = [d.\mathbf{1}.\mathbf{1}' + (1-d)\mathbf{I}]$ para encontrar los componentes y discutir su interpretación).

5.3. Supongamos que $Z, X_1, ..., X_p$ tienen una distribución normal $(p+1)$ dimensional. Sean $Y_1, ..., Y_p$ los componentes principales de $X_1, ..., X_p$. Demostrar que el coeficiente de correlación múltiple de las regresiones:

$$Z = \sum a_i X_i$$
$$Z = \sum b_i Y_i$$

es idéntico.

5.4. Demostrar que si $\mathbf{S} = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}$, donde $\mathbf{A}$ y $\mathbf{B}$ son no singulares de rango r_A y r_B los vectores propios de $\mathbf{S}$ son de la forma $(\mathbf{u}_1, 0)$ y $(0, \mathbf{u}_2)$, donde $\mathbf{u}_1$ es un vector propio de $\mathbf{A}$ y $\mathbf{u}_2$ un vector propio de $\mathbf{B}$.

5.5. Indicar las implicaciones del resultado del Ejercicio 5.4 para calcular componentes principales.

5.6. Demostrar que si $\mathbf{S} = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}$ los valores propios de $\mathbf{S}$ son los de $\mathbf{A}$ más los de $\mathbf{B}$.

5.7. Demostrar que el espacio que maximiza la varianza generalizada de la proyección es el definido por $z_1 = \mathbf{X}\mathbf{a}_1$ y $z_2 = \mathbf{X}\mathbf{a}_2$ donde z_1 y z_2 son los dos primeros componentes principales.

5.8. Demostrar que si una variable x_1 está incorrelada con el resto de manera que la matriz $\mathbf{S}$ tiene la forma $\mathbf{S} = \begin{bmatrix} s_1^2 & \mathbf{0}' \\ \mathbf{0} & \mathbf{S_2} \end{bmatrix}$ donde $\mathbf{0}$ y $\mathbf{0}'$ son vectores de ceros, la matriz $\mathbf{S}$ tiene un componente principal asociado únicamente a la primera variable, es decir, el vector (1, 0,...,0) es un vector propio de $\mathbf{S}$.

5.9. Demostrar que la dirección donde la variabilidad de la proyección es mínima es la dada por el vector propio ligado al menor valor propio de la matriz de covarianzas.

5.10. Demostrar la siguiente acotación para formas cuadráticas: $\lambda_{\min} w'w \leq w'\mathbf{B}w \leq \lambda_{\max} \leq w'w$, donde $\lambda_{\min}$ y $\lambda_{\max}$ son el menor y el mayor valor propio de la matriz $\mathbf{B}$. (Sugerencia, maximizar la forma cuadrática como se hizo para obtener el primer componente principal.)

APÉNDICE 5.1.
DISTANCIAS ENTRE PUNTOS Y PROYECCIONES

Vamos a demostrar que maximizar las distancias al cuadrado entre los puntos proyectados equivale a maximizar la varianza de la variable definida por las proyecciones de los puntos. Sea $z_i = \mathbf{a}_1'\mathbf{x}_i$ la proyección de una observación sobre la dirección $\mathbf{a}_1$, donde suponemos $\mathbf{a}_1'\mathbf{a}_1 = 1$. La variable z_i tendrá media cero ya que si las x tienen media cero $\sum_{i=1}^n z_i = \sum_{i=1}^n \mathbf{a}_1'\mathbf{x}_i = = \mathbf{a}_1' \sum_{i=1}^n \mathbf{x}_i = 0$. La suma de las distancias al cuadrado entre los puntos proyectados es

$$D_p = \sum_{i=1}^{n} \sum_{h=i+1}^{n} (z_i - z_h)^2.$$

Para interpretar este sumatorio observemos que cada término z_i aparece al cuadrado $n-1$, veces ya que cada punto se compara con los otros $n-1$, y que habrá tantos dobles productos como parejas de puntos, es decir $\binom{n}{2} = n(n-1)/2$. Por tanto:

$$D_p = (n-1)\sum_{i=1}^{n} z_i^2 - 2\sum_{i=1}^{n}\sum_{h=i+1}^{n} z_i z_h = n\sum z_i^2 - B$$

siendo B:

$$B = \sum_{i=1}^{n} z_i^2 + 2\sum_{i=1}^{n}\sum_{h=i+1}^{n} z_i z_h$$

que puede escribirse,

$$\begin{aligned} B &= z_1(z_1 + z_2 + \ldots z_n) + z_2(z_1 + \ldots + z_n) + \ldots z_n(z_1 + \ldots + z_n) \\ &= \sum_{i=1}^{n} z_i \sum_{i=1}^{n} z_i = 0. \end{aligned}$$

Por tanto, maximizar las distancias entre los puntos equivale a maximizar:

$$A = n\sum z_i^2$$

que es el criterio de maximizar la varianza de la nueva variable, obtenida anteriormente.

Algunos autores han propuesto minimizar

$$\sum\sum w_{ij}(d_{ij} - \widehat{d}_{ij})^2$$

donde w_{ij} es una función de ponderación. El problema así planteado no tiene una solución simple y debe resolverse mediante un algoritmo iterativo no lineal. Veáse, por ejemplo, Krzanowski (1990, Cap. 2).

APÉNDICE 5.2.
LOS COMPONENTES COMO PREDICTORES ÓPTIMOS

Demostraremos que los componentes principales son predictores óptimos de las $\mathbf{X}$. Comencemos demostrando que si queremos aproximar la matriz $\mathbf{X}$, de rango p, por otra matriz $\widehat{\mathbf{X}}_r$ de rango $r < p$, la aproximación óptima es $\hat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r'$ donde la matriz $\mathbf{A}_r$ es $p \times r$ y sus columnas son los vectores propios asociados a los r mayores valores propios de la matriz $\mathbf{S}$.

El problema de aproximar la matriz $\mathbf{X}$ puede establecerse así: consideremos un espacio de dimensión r definido por una base $\mathbf{U}_r$ ortonormal, donde $\mathbf{U}_r$ es $p \times r$ y $\mathbf{U}_r'\mathbf{U}_r = \mathbf{I}$. Se desea encontrar una aproximación de la matriz $\mathbf{X}$ utilizando esta base, es decir, queremos prever cada una de las filas $(\mathbf{x}_1, ..., \mathbf{x}_n)$ de la matriz, donde $\mathbf{x}_i$ es el vector $p \times 1$ de observaciones en el elemento i de la muestra, mediante los vectores $\mathbf{U}_r$. La predicción de la variable $\mathbf{x}_i$ será la proyección ortogonal sobre el espacio generado por estos vectores (véase la sección 2.5), que es:

$$\widehat{\mathbf{x}}_i = \mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i \tag{5.17}$$

y queremos determinar los vectores $\mathbf{U}_r$ tal que el error cuadrático de estas predicciones sea mínimo. El error cuadrático para todos los elementos de la matriz $\mathbf{X}$ viene dado por:

$$E = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \widehat{x}_{ij})^2 = \sum_{i=1}^{n}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)'(\mathbf{x}_i - \widehat{\mathbf{x}}_i)$$

y queremos que sea mínimo. El error puede escribirse, utilizando (5.17).

$$E = \sum_{i=1}^{n}\mathbf{x}_i{}'\mathbf{x}_i - \sum_{i=1}^{n}\mathbf{x}_i{}'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i \tag{5.18}$$

Minimizar el error equivale a maximizar el segundo término. Utilizando que un escalar es igual a su traza, $\sum_{i=1}^{n} \mathbf{x}_i{}'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i = tr(\sum_{i=1}^{n} \mathbf{x}_i{}'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i) = \sum_{i=1}^{n} tr(\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i\mathbf{x}_i{}') = = tr(\mathbf{U}_r\mathbf{U}_r' \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}')$. Introduciendo que $\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}'/n$ y sustituyendo en $tr(\mathbf{U}_r\mathbf{U}_r' \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}')$, tenemos que esta expresión es $ntr(\mathbf{U}_r\mathbf{U}_r'\mathbf{S}) = ntr(\mathbf{U}_r'\mathbf{S}\mathbf{U}_r)$. Por tanto: $\sum_{i=1}^{n} \mathbf{x}_i{}' \mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i = ntr(\mathbf{U}_r'\mathbf{S}\mathbf{U}_r)$.

Según esta expresión, minimizar el error (5.18) implica encontrar un conjunto de vectores $\mathbf{U}_r = [\mathbf{u}_1, ..., \mathbf{u}_r]$ que maximicen la suma de los elementos diagonales de $\mathbf{U}_r'\mathbf{S}\mathbf{U}_r$, es decir, maximicen $\sum_{j=1}^{r} \mathbf{u}_j'\mathbf{S}\mathbf{u}_j$. Si $r = 1$, éste es el problema que se ha resuelto para encontrar el primer componente. Si $r = 2$, como el nuevo vector debe ser ortogonal al primero, obtenemos el segundo componente, y así sucesivamente. Por tanto, $\mathbf{U}_r = \mathbf{A}_r$, y la aproximación óptima a la matriz $\mathbf{X}$ vendrá dada por $\widehat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r'$. Además, como en (5.18) el primer término es:

$$\sum_{i=1}^{n} \mathbf{x}_i{}'\mathbf{x}_i = tr(\sum_{i=1}^{n} \mathbf{x}_i{}'\mathbf{x}_i) = \sum_{i=1}^{n} tr(\mathbf{x}_i'\mathbf{x}_i) =$$

$$= tr \sum_{i=1}^{n} (\mathbf{x}_i\mathbf{x}_i{}') = ntr(\mathbf{S}) = n \sum_{i=1}^{p} \lambda_i$$

y el segundo es, según (5.19), igual a $n \sum_{i=1}^{r} \lambda_i$, tenemos que el error de la aproximación será $n \sum_{i=r+1}^{p} \lambda_i$.

Es interesante señalar que esta aproximación es la que proporciona la descomposición en valores singulares (véase la Sección 2.4.5), es decir, la mejor aproximación a la matriz $\mathbf{X}$ por otra matriz $\widehat{\mathbf{X}}_r$ de rango $r < p$ es:

$$\widehat{\mathbf{X}}_r = \mathbf{U}_r\mathbf{D}_r^{1/2}\mathbf{V}_r' = \sum_{i=1}^{r} \lambda_1^{1/2}\mathbf{u}_i\mathbf{v}_i'$$

donde $\mathbf{U}_r$ es la matriz de los r mayores vectores propios de $\mathbf{X}\mathbf{X}'$, $\mathbf{D}_r^{1/2}$ contiene los r mayores valores propios y $\mathbf{V}_r$ contiene los r vectores propios de $\mathbf{X}'\mathbf{X}$·. En efecto, como $\mathbf{A}_r = \mathbf{V}_r$, $\hat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r' = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{V}'\mathbf{V}_r\mathbf{V}_r' = \mathbf{U}_r\mathbf{D}_r^{\frac{1}{2}}\mathbf{V}_r'$. Además observemos que la solución $\hat{\mathbf{X}}_r$ es el resultado de proyectar las filas de $\mathbf{X}$ sobre el espacio definido por los r vectores propios ligados a los r mayores valores propios de $\mathbf{X}'\mathbf{X}$, ya que $\hat{\mathbf{X}}_r' = \mathbf{A}_r\mathbf{A}_r'\mathbf{X}'$, que es una proyección ortogonal.

El problema puede enfocarse desde otro punto de vista: buscar unas variables $[\mathbf{z}_1, ..., \mathbf{z}_r]$ que sean combinaciones lineales de las originales y que tengan la propiedad de preverlas de manera óptima. Por ejemplo, si $r = 1$, buscamos un vector $\mathbf{a}_1$ de manera que la nueva variable:

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$$

permita prever con mínimo error los valores observados para el conjunto de variables que forman las columnas de la matriz $\mathbf{X}$. Por ejemplo, el valor previsto para la variable x_j en el individuo i, $\widehat{x}_{ij}$, conocido el valor de la variable z_1 para ese individuo, z_{1i} será:

$$\widehat{x}_{ij} = b_j z_{1i}$$

y el error de predicción será $\mathrm{e}_{ij} = x_{ij} - \widehat{x}_{ij}$. Vamos a demostrarlo, para simplificar, en el caso $r = 1$. Calcularemos el vector $\mathbf{a}_1$ que minimiza estos errores de predicción. El coeficiente de regresión b_j viene dado por:

$$b_j = \frac{\sum_{i=1}^{n} x_{ij}z_{1i}}{\sum z_{1i}^2} \tag{5.19}$$

como $1/n \sum z_{1i}^2 = (1/n)\mathbf{a'X'Xa} = \mathbf{a'Sa}$, la varianza de z_1 puede crecer indefinidamente si no imponemos ninguna restricción. Exigiremos que sea unitaria, es decir que:

$$\mathbf{a'Sa} = 1 = (1/n) \sum z_{1i}^2 \tag{5.20}$$

Entonces:

$$b_j = 1/n \sum x_{ij} z_{1i} = (1/n)\mathbf{X}_j'\mathbf{X}\mathbf{a}_1 = \mathbf{V}_j'\mathbf{a}_1 \tag{5.21}$$

donde $\mathbf{V}_j$ es el vector fila j de la matriz $\mathbf{S}$ de varianzas y covarianzas. Impongamos la condición mínimo cuadrática para obtener $\mathbf{a_1}$:

$$\frac{1}{n}\sum_{i=1}^{n} e_{ij}^2 = \text{Mínimo} = \frac{1}{n}\sum_{i=1}^{n} \left(x_{ij} - \mathbf{V}_j'\mathbf{a}_1 z_{1i}\right)^2$$

y el segundo miembro puede escribirse:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + \frac{1}{n}\mathbf{a}_1'\mathbf{V}_j\mathbf{V}_j'\mathbf{a}_1 \sum_{i=1}^{n} z_{1i}^2 - 2\mathbf{V}_j'\mathbf{a}\frac{1}{n}\sum_{i=1}^{n} x_{ij} z_{1i}$$

utilizando ahora (5.20) y (5.21), se obtiene

$$\frac{1}{n}\sum_{i=1}^{n} e_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 - \mathbf{a}_1'\mathbf{V}_j\mathbf{V}_j'\mathbf{a}_1.$$

Aplicando este mismo razonamiento a las otras variables X y sumando para todas ellas:

$$M = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} e_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 - \sum_{j=1}^{p} \mathbf{a}_1'\mathbf{V}_j\mathbf{V}_j'\mathbf{a}_1$$

como el primer miembro es la traza de $\mathbf{S}$ que es fija, maximizar M equivale a minimizar:

$$\mathbf{a}_1' \sum_{j=1}^{p} \mathbf{V}_j\mathbf{V}_j'\mathbf{a}_1 = \mathbf{a}_1'\mathbf{S}\mathbf{S}'\mathbf{a}_1' = \mathbf{a}_1'\mathbf{S}^2\mathbf{a}_1 \tag{5.22}$$

ya que $\mathbf{S}$ es simétrica. Por tanto, el problema es minimizar la expresión (5.22) con la restricción (5.20):

$$L = \mathbf{a}_1'\mathbf{S}^2\mathbf{a}_1 - \lambda(\mathbf{a}_1\mathbf{S}\mathbf{a}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{a}} = 2\mathbf{S}^2\mathbf{a} - \lambda 2\mathbf{S}\mathbf{a} = 0$$

$$\mathbf{S}^2\mathbf{a} = \lambda\mathbf{S}\mathbf{a}$$

de donde intuimos que $\mathbf{a}$ debe de ser un vector propio de $\mathbf{S}$ y λ un valor propio, ya que si:

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$$

multiplicando por $\mathbf{S}$

$$\mathbf{S}^2\mathbf{a} = \lambda\mathbf{S}\mathbf{a}$$

Con lo que finaliza la demostración.