

Análisis de Componentes Principales

Basado en las notas de Norberto Corral Blanco y Beatriz Sinova Fernández
Facultad de Ciencias, Universidad de Oviedo

Índice

1. Introducción	2
2. Planteamiento y primeras consideraciones	2
3. Primera componente principal	2
3.1. Maximización de la varianza	2
3.2. Solución mediante descomposición espectral	3
3.3. Unicidad y signo	4
4. Segunda componente principal	4
5. Todas las componentes principales	5
6. Caso de matriz de covarianzas singular	5
7. Elección del número de componentes	5
7.1. Variabilidad explicada	5
7.2. Principio de imparcialidad	6
7.3. Variable incorrelada con el resto	6
8. Calidad de la reconstrucción de las variables originales	6
9. Interpretación de las componentes mediante correlaciones	7
9.1. Perfil de individuos extremos	7
9.2. Correlaciones entre variables originales y componentes	7
9.3. Observación sobre el signo	7
10. Componentes principales normadas (estandarización previa)	7
11. Componentes principales sobre una muestra	8
11.1. Notación y matriz de datos	8
11.2. Centrado de los datos	8
11.3. La matriz de centrado	8
11.4. Matriz de varianzas-covarianzas muestral	9
11.5. Cálculo de las componentes principales muestrales	9
11.6. Variabilidad explicada y elección del número de componentes	10

1. Introducción

El análisis de componentes principales (ACP) es un método de reducción de la dimensión cuyo objetivo fundamental es representar datos originales multidimensionales en un espacio de menor dimensión, conservando en la medida de lo posible la estructura inicial o esencia de los datos (por ejemplo, la distancia entre los puntos). En otras palabras, se busca describir de la forma más sencilla posible la estructura de una población que cuenta con muchas variables, mediante un número reducido de características.

La técnica fue introducida por Hotelling (1933), aunque sus orígenes se remontan a los ajustes ortogonales por mínimos cuadrados propuestos por K. Pearson (1901).

Para ilustrar la idea intuitiva, se puede pensar en la representación de un bolígrafo tridimensional en un plano bidimensional. Las posiciones relativas de los puntos proyectados en el plano deben ser lo más parecidas posible a las posiciones originales en el espacio; es decir, la proyección ha de hacerse de modo que los puntos proyectados conserven la distancia original entre ellos, o lo que es equivalente, que conserven al máximo la variabilidad de los datos originales.

Nos centraremos en un vector aleatorio $\vec{x} = (x_1, x_2, \dots, x_p)^t$ del que queremos obtener una representación simplificada $\vec{y} = g(\vec{x})$ con menor dimensión. Interesa preservar la estructura de los datos; en concreto, queremos conservar su variabilidad, ya que las medidas de dispersión y de relación entre variables (recogidas en la matriz de varianzas-covarianzas) caracterizan gran parte de la información contenida en \vec{x} .

2. Planteamiento y primeras consideraciones

La propuesta más sencilla es utilizar transformaciones lineales de las variables originales. Trabajaremos con las variables centradas, ya que la variabilidad del objeto es independiente de su posición (recogida en el vector de medias $\vec{\mu} = E(\vec{x})$). Por tanto, definimos

$$y = \vec{a}^t(\vec{x} - \vec{\mu}), \quad (1)$$

donde $\vec{a} = (a_1, \dots, a_p)^t$ es un vector de coeficientes que determina la combinación lineal. El objetivo será encontrar \vec{a} de manera que y conserve el máximo de variabilidad.

3. Primera componente principal

3.1. Maximización de la varianza

Queremos que el objeto proyectado se asemeje lo más posible al objeto real, por lo que buscamos maximizar la dispersión de la nueva variable:

$$\max_{\vec{a}} \text{Var}(y).$$

Dado que las variables están centradas, $E(y) = 0$:

$$E(y) = E(\vec{a}^t(\vec{x} - \vec{\mu})) = \vec{a}^t E(\vec{x} - \vec{\mu}) = \vec{a}^t(\vec{\mu} - \vec{\mu}) = 0.$$

Por tanto,

$$\text{Var}(y) = E(y^2) = E(\vec{a}^t(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^t \vec{a}) = \vec{a}^t E((\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^t) \vec{a} = \vec{a}^t \Sigma \vec{a},$$

donde $\Sigma = \text{Cov}(\vec{x})$ es la matriz de varianzas-covarianzas de \vec{x} , que es simétrica y semidefinida positiva.

Si no se impone ninguna restricción sobre \vec{a} , el problema no tiene solución, ya que dado cualquier \vec{a}_0 que proporcione una varianza positiva, el vector $\vec{a}_1 = k\vec{a}_0$ con $k > 1$ satisface

$$\vec{a}_1^t \Sigma \vec{a}_1 = k^2 \vec{a}_0^t \Sigma \vec{a}_0 > \vec{a}_0^t \Sigma \vec{a}_0,$$

y haciendo crecer k la varianza se puede hacer arbitrariamente grande. Por ello, es necesario limitar la norma del vector de coeficientes. La restricción habitual es $\|\vec{a}\| = 1$.

3.2. Solución mediante descomposición espectral

Planteamos:

$$\max_{\|\vec{a}\|=1} \vec{a}^t \Sigma \vec{a}.$$

Dado que Σ es simétrica y semidefinida positiva, admite una descomposición espectral

$$\Sigma = U \Lambda U^t,$$

donde:

- $U = (\vec{u}_1 \ \vec{u}_2 \ \cdots \ \vec{u}_p)$ es una matriz ortogonal ($U^t U = U U^t = I$), cuyas columnas son los vectores propios ortonormales de Σ .
- $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ es la matriz diagonal de valores propios, ordenados sin pérdida de generalidad de mayor a menor: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Definamos $\vec{b} = U^t \vec{a}$. Como U es ortogonal, la transformación $\vec{a} \mapsto \vec{b}$ es una biyección que conserva la norma:

$$\|\vec{b}\|^2 = \vec{b}^t \vec{b} = \vec{a}^t U U^t \vec{a} = \vec{a}^t \vec{a} = \|\vec{a}\|^2.$$

Entonces,

$$\vec{a}^t \Sigma \vec{a} = \vec{a}^t U \Lambda U^t \vec{a} = \vec{b}^t \Lambda \vec{b} = \sum_{j=1}^p \lambda_j b_j^2.$$

El problema original es equivalente a

$$\max_{\|\vec{b}\|=1} \sum_{j=1}^p \lambda_j b_j^2.$$

Acotando,

$$\sum_{j=1}^p \lambda_j b_j^2 \leq \lambda_1 \sum_{j=1}^p b_j^2 = \lambda_1, \quad \text{pues } \sum_{j=1}^p b_j^2 = \|\vec{b}\|^2 = 1.$$

La cota superior λ_1 se alcanza eligiendo $\vec{b} = \vec{e}_1 = (1, 0, \dots, 0)^t$, lo que da $\sum \lambda_j b_j^2 = \lambda_1$. Deshaciendo el cambio,

$$\vec{a} = U \vec{b} = U \vec{e}_1 = \vec{u}_1.$$

Por tanto, el vector de coeficientes que maximiza la varianza es el primer vector propio \vec{u}_1 , y la varianza máxima alcanzada es λ_1 :

$\max_{\ \vec{a}\ =1} \text{Var}(y) = \lambda_1, \quad \text{con } \vec{a} = \vec{u}_1.$
--

La variable obtenida $y_1 = \vec{u}_1^t (\vec{x} - \vec{\mu})$ se denomina **primera componente principal**.

3.3. Unicidad y signo

La solución no es única en sentido estricto. Si \vec{u}_1 es un vector propio normalizado, $-\vec{u}_1$ también lo es. Por ello, a la hora de interpretar los coeficientes no debe prestarse atención al signo de forma aislada, sino a su relación con el signo del resto de los coeficientes. Además, si el valor propio λ_1 tiene multiplicidad geométrica mayor que uno, existirán infinitas soluciones (todo un subespacio) que generan componentes igualmente válidas; en tal caso la elección de una base concreta deberá hacerse atendiendo a criterios de interpretabilidad adicionales.

4. Segunda componente principal

Para seguir extrayendo información no recogida por la primera componente, busquemos una segunda variable $y_2 = \vec{a}_2^t(\vec{x} - \vec{\mu})$ que:

1. Maximice su varianza: $\max_{\vec{a}_2} \vec{a}_2^t \Sigma \vec{a}_2$,
2. Tenga norma unitaria: $\|\vec{a}_2\| = 1$,
3. Aporte información nueva, es decir, que no esté correlacionada con la primera componente: $\text{Cov}(y_1, y_2) = 0$.

Como ambas variables están centradas,

$$\text{Cov}(y_1, y_2) = E(y_1 y_2) = E(\vec{u}_1^t(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^t \vec{a}_2) = \vec{u}_1^t \Sigma \vec{a}_2.$$

Puesto que \vec{u}_1 es vector propio de Σ con valor propio λ_1 , $\Sigma \vec{u}_1 = \lambda_1 \vec{u}_1$, y teniendo en cuenta la simetría,

$$\vec{u}_1^t \Sigma = \lambda_1 \vec{u}_1^t.$$

Por tanto,

$$\text{Cov}(y_1, y_2) = \lambda_1 \vec{u}_1^t \vec{a}_2.$$

Salvo el caso degenerado $\lambda_1 = 0$ (que implicaría varianza nula de la primera componente), la condición $\text{Cov}(y_1, y_2) = 0$ equivale a la ortogonalidad

$$\vec{u}_1^t \vec{a}_2 = 0. \quad (2)$$

Sea de nuevo $\vec{b}_2 = U^t \vec{a}_2$. Entonces $\|\vec{b}_2\| = \|\vec{a}_2\| = 1$, y la condición (2) se traduce en que la primera coordenada de \vec{b}_2 es nula:

$$b_{21} = \vec{u}_1^t \vec{a}_2 = 0.$$

Así,

$$\vec{a}_2^t \Sigma \vec{a}_2 = \vec{b}_2^t \Lambda \vec{b}_2 = \sum_{j=2}^p \lambda_j b_{2j}^2,$$

pues $b_{21} = 0$. Acotando,

$$\sum_{j=2}^p \lambda_j b_{2j}^2 \leq \lambda_2 \sum_{j=2}^p b_{2j}^2 \leq \lambda_2 \sum_{j=1}^p b_{2j}^2 = \lambda_2.$$

Esta cota se alcanza con $\vec{b}_2 = \vec{e}_2 = (0, 1, 0, \dots, 0)^t$, que cumple $b_{21} = 0$ y $\|\vec{b}_2\| = 1$. Por tanto, el vector óptimo es $\vec{a}_2 = U \vec{e}_2 = \vec{u}_2$, y la varianza máxima es λ_2 . Así,

$$\boxed{y_2 = \vec{u}_2^t(\vec{x} - \vec{\mu}), \quad \text{Var}(y_2) = \lambda_2, \quad \text{Cov}(y_1, y_2) = 0.}$$

5. Todas las componentes principales

Este razonamiento se extiende de forma inductiva. En el paso k -ésimo ($k \leq p$) se busca

$$y_k = \vec{a}_k^t(\vec{x} - \vec{\mu})$$

con $\|\vec{a}_k\| = 1$, que maximice $\text{Var}(y_k)$ sujeta a las condiciones de incorrelación con las $k-1$ componentes anteriores:

$$\text{Cov}(y_k, y_j) = 0 \quad \forall j < k \iff \vec{u}_j^t \vec{a}_k = 0 \quad \forall j < k.$$

Repitiendo el proceso, la solución es $\vec{a}_k = \vec{u}_k$, el k -ésimo vector propio de Σ , y la varianza de y_k es λ_k . De este modo, las p componentes principales (supuesto que Σ tiene rango completo) vienen dadas por

$$\boxed{\vec{y} = U^t(\vec{x} - \vec{\mu}),}$$

donde $y_j = \vec{u}_j^t(\vec{x} - \vec{\mu})$, con $\text{Var}(y_j) = \lambda_j$ y $\text{Cov}(y_i, y_j) = 0$ para $i \neq j$.

6. Caso de matriz de covarianzas singular

Si Σ es singular, su rango q es menor que p , y existirán $p - q$ valores propios nulos: $\lambda_{q+1} = \dots = \lambda_p = 0$. Para estos,

$$\text{Var}(y_j) = \lambda_j = 0,$$

y por estar centradas, $y_j \equiv 0$ (casi seguramente). Estas componentes no aportan información y pueden ser omitidas. En consecuencia, con q componentes principales se puede representar exactamente toda la variabilidad del vector \vec{x} . Desde el punto de vista práctico, ello indica que hay redundancia lineal en las variables originales (alguna variable es combinación lineal exacta de las demás) y el problema se reduce a un espacio de dimensión q .

7. Elección del número de componentes

En la práctica, se retienen solo unas pocas componentes que capturen la mayor parte de la variabilidad. Para decidir cuántas, se utilizan los siguientes criterios.

7.1. Variabilidad explicada

La variabilidad total de las p variables originales se puede medir por la traza de Σ :

$$\text{Var}(\text{total}) = \text{tr}(\Sigma) = \sum_{i=1}^p \lambda_i.$$

La proporción de variabilidad explicada por la j -ésima componente es

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

Por consiguiente, las q primeras componentes explican una fracción

$$\frac{\sum_{j=1}^q \lambda_j}{\sum_{i=1}^p \lambda_i}$$

de la variabilidad total. Como guía orientativa, se suele buscar un número q que explique alrededor del 70 %–80 % de la variabilidad total, aunque el equilibrio entre simplificación (pocas componentes) y fidelidad (alto porcentaje) es una decisión del analista.

7.2. Principio de imparcialidad

Cuando varios valores propios son muy parecidos (incluyendo el caso de multiplicidad algebraica r), se recomienda retener todos o ninguno, para evitar favorecer arbitrariamente una dirección frente a otras equivalentes. Por ejemplo, si los porcentajes explicados por las componentes son 40 %, 25 % y 25 %, se podrían elegir las tres (para alcanzar un 90 % acumulado) o, si no se alcanza un umbral suficiente, renunciar también a las dos del 25 %. La decisión ha de justificarse en cada caso.

7.3. Variable incorrelada con el resto

Si una variable original x_1 está incorrelada con las demás, la matriz de covarianzas toma la forma

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \mathbf{0}^t \\ \mathbf{0} & \Sigma_2 \end{pmatrix}.$$

Un sencillo cálculo muestra que

$$\vec{u}_1 = (1, 0, \dots, 0)^t$$

es vector propio de Σ asociado al valor propio σ_1^2 . Por tanto, existe una componente principal que es (salvo factor de escala) la propia variable x_1 centrada. Si σ_1^2 es grande, esta componente aparecerá entre las primeras; si es pequeño, entre las últimas. Esta propiedad ayuda a interpretar las componentes: una variable muy independiente del resto puede manifestarse como una componente aislada.

8. Calidad de la reconstrucción de las variables originales

Una vez seleccionadas q componentes, podemos preguntarnos qué parte de la variabilidad de cada variable original x_i somos capaces de recuperar. Usando la relación $\vec{x} - \vec{\mu} = U\vec{y}$, tenemos

$$x_i - \mu_i = \sum_{j=1}^p u_{ij} y_j = \sum_{j=1}^q u_{ij} y_j + \sum_{j=q+1}^p u_{ij} y_j,$$

donde la primera suma corresponde a la parte reconstruida con las q componentes y la segunda es la parte que se descarta. Como las componentes y_j son incorreladas y $\text{Var}(y_j) = \lambda_j$, la varianza de la parte reconstruida es

$$\text{Var}\left(\sum_{j=1}^q u_{ij} y_j\right) = \sum_{j=1}^q u_{ij}^2 \lambda_j.$$

Así, la proporción de variabilidad de x_i explicada por las q componentes es

$$\frac{\sum_{j=1}^q u_{ij}^2 \lambda_j}{\Sigma_{ii}}.$$

Este cociente indica la calidad de la reconstrucción de cada variable. Si para cierta variable es muy bajo, conviene examinar si las componentes seleccionadas la representan deficientemente.

9. Interpretación de las componentes mediante correlaciones

9.1. Perfil de individuos extremos

Un primer acercamiento consiste en examinar los individuos con puntuaciones extremas en cada componente e intentar caracterizar qué variables los distinguen del resto.

9.2. Correlaciones entre variables originales y componentes

Una herramienta más sistemática es el cálculo de las correlaciones entre cada variable original x_i y cada componente y_j . Como $\vec{x} - \vec{\mu} = U\vec{y}$,

$$\text{Cov}(\vec{x}, \vec{y}) = E((\vec{x} - \vec{\mu}) \vec{y}^t) = E(U\vec{y} \vec{y}^t) = U E(\vec{y} \vec{y}^t) = U\Lambda.$$

Por tanto,

$$\text{Cov}(x_i, y_j) = \lambda_j u_{ij}.$$

Recordando que $\text{Var}(y_j) = \lambda_j$ y denotando $\text{Var}(x_i) = \Sigma_{ii}$, la correlación es

$$\text{Corr}(x_i, y_j) = \frac{\lambda_j u_{ij}}{\sqrt{\Sigma_{ii} \lambda_j}} = \frac{\sqrt{\lambda_j} u_{ij}}{\sqrt{\Sigma_{ii}}}.$$

Valores absolutos próximos a 1 indican que la variable x_i y la componente y_j miden esencialmente el mismo rasgo; valores cercanos a 0 revelan una influencia despreciable. De esta manera, observando qué variables originales presentan altas correlaciones (en valor absoluto) con una componente, se puede poner una “etiqueta” semántica a la componente.

9.3. Observación sobre el signo

Al igual que ocurre con los coeficientes de los vectores propios, el signo de las correlaciones no es interpretable de manera absoluta, porque cada vector propio puede cambiarse de signo sin afectar a las propiedades estadísticas. Lo relevante es el signo relativo entre las distintas variables que se asocian a una misma componente.

10. Componentes principales normadas (estandarización previa)

Cuando las variables originales se miden en unidades muy diferentes (por ejemplo, kilogramos frente a metros, o escalas con distinta dispersión), las que presentan magnitudes numéricas mayores tenderán a dominar el análisis, ya que los valores de la matriz Σ dependen de las escalas. Para evitar este efecto, se recomienda estandarizar las variables (restar la media y dividir por la desviación típica) antes de extraer las componentes. Equivalentemente, se trabaja con la matriz de correlaciones R en lugar de la matriz de covarianzas Σ :

$$R = \text{diag}(\Sigma)^{-1/2} \Sigma \text{diag}(\Sigma)^{-1/2}.$$

Las componentes principales normadas se obtienen como los vectores propios de R (ordenados según sus valores propios), y equivalen a aplicar el ACP a las variables tipificadas. La interpretación y los criterios de selección son análogos, pero con la ventaja de que todas las variables contribuyen en igualdad de condiciones (ahora $\text{tr}(R) = p$).

11. Componentes principales sobre una muestra

11.1. Notación y matriz de datos

En la práctica no se conoce la matriz de covarianzas poblacional Σ , sino que se dispone de una muestra de n individuos en los que se miden p variables. La información se organiza en una matriz de datos X de dimensión $n \times p$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Donde la fila i representa el vector de p características del individuo i , es decir $\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$, y la columna j el vector de n valores de la variable j -ésima, $\vec{x}_{(j)} = (x_{1j}, x_{2j}, \dots, x_{nj})^t$.

11.2. Centrado de los datos

Para eliminar el efecto de la localización trabajamos con los datos centrados. Sea $\mathbf{1}_n = (1, 1, \dots, 1)^t$ el vector de n unos y $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ la media de la variable j . La columna centrada correspondiente es

$$\vec{x}_{c(j)} = \begin{pmatrix} x_{1j} - \bar{x}_j \\ x_{2j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{pmatrix} = \vec{x}_{(j)} - \bar{x}_j \mathbf{1}_n.$$

En forma matricial, la matriz de datos centrados X_c (de tamaño $n \times p$) será

$$X_c = (\vec{x}_{c(1)} \vec{x}_{c(2)} \cdots \vec{x}_{c(p)}).$$

11.3. La matriz de centrado

Para unificar notación, se introduce la matriz de centrado de tamaño $n \times n$:

$$H = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t,$$

donde $J_n = \mathbf{1}_n \mathbf{1}_n^t$ es la matriz cuadrada de unos (todas sus entradas son 1). Nótese que $\mathbf{1}_n^t \mathbf{1}_n = n$, por lo que también podemos escribir

$$H = I_n - \mathbf{1}_n (\mathbf{1}_n^t \mathbf{1}_n)^{-1} \mathbf{1}_n^t.$$

Propiedades fundamentales de H :

1. **Simetría:** $H^t = H$. En efecto, $I_n^t = I_n$ y $(\mathbf{1}_n \mathbf{1}_n^t)^t = \mathbf{1}_n \mathbf{1}_n^t$.

2. **Idempotencia:** $H^2 = H$. Multiplicando,

$$H^2 = \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t\right)^2 = I_n - \frac{2}{n} \mathbf{1}_n \mathbf{1}_n^t + \frac{1}{n^2} \mathbf{1}_n \mathbf{1}_n^t \mathbf{1}_n \mathbf{1}_n^t = I_n - \frac{2}{n} \mathbf{1}_n \mathbf{1}_n^t + \frac{n}{n^2} \mathbf{1}_n \mathbf{1}_n^t = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t = H.$$

3. **Valores propios:** Al ser idempotente, sus valores propios solo pueden ser 0 o 1. El vector $\mathbf{1}_n$ es vector propio con valor propio 0, pues

$$H \mathbf{1}_n = \mathbf{1}_n - \frac{1}{n} \mathbf{1}_n (\mathbf{1}_n^t \mathbf{1}_n) = \mathbf{1}_n - \frac{n}{n} \mathbf{1}_n = \mathbf{0}.$$

Cualquier vector ortogonal a $\mathbf{1}_n$ es vector propio con valor propio 1. Por tanto, $\text{rango}(H) = n - 1$ y $\text{tr}(H) = n - 1$.

La matriz de centrado actúa sobre X produciendo los datos centrados:

$$X_c = HX.$$

En efecto, la columna j de HX es

$$H \vec{x}_{(j)} = \vec{x}_{(j)} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t \vec{x}_{(j)} = \vec{x}_{(j)} - \bar{x}_j \mathbf{1}_n = \vec{x}_{c(j)}.$$

11.4. Matriz de varianzas-covarianzas muestral

La matriz de varianzas-covarianzas muestral $S = (s_{ij})_{p \times p}$ contiene en la diagonal las varianzas muestrales

$$s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2,$$

y fuera de ella, las covarianzas muestrales

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

En forma matricial compacta:

$$S = \frac{1}{n} X_c^t X_c = \frac{1}{n} (HX)^t (HX) = \frac{1}{n} X^t H^t H X.$$

Usando la simetría e idempotencia de H ($H^t H = H^2 = H$), se obtiene la expresión más habitual:

$$\boxed{S = \frac{1}{n} X^t H X.}$$

Nótese que muchos textos dividen por $n - 1$ en lugar de n ; en ese caso se reemplaza $\frac{1}{n}$ por $\frac{1}{n-1}$ y todas las fórmulas siguientes se modifican de manera análoga. Con la variante $\frac{1}{n}$ la matriz S es un estimador máximo-verosímil bajo normalidad.

11.5. Cálculo de las componentes principales muestrales

Queremos encontrar combinaciones lineales de las variables centradas

$$\vec{y} = X_c \vec{a} = HX\vec{a}$$

(donde \vec{a} es un vector de coeficientes $p \times 1$ con $\|\vec{a}\| = 1$) que maximicen la varianza muestral. La media de \vec{y} es nula por construcción, y su varianza muestral es

$$s_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 = \frac{1}{n} \vec{y}^t \vec{y} = \frac{1}{n} \vec{a}^t X^t H^t H X \vec{a} = \vec{a}^t \left(\frac{1}{n} X^t H X \right) \vec{a} = \vec{a}^t S \vec{a}.$$

El problema de maximización es, por tanto, formalmente idéntico al poblacional:

$$\max_{\|\vec{a}\|=1} \vec{a}^t S \vec{a}.$$

Como S es una matriz simétrica y semidefinida positiva, admite una descomposición espectral

$$S = \widehat{U} \widehat{\Lambda} \widehat{U}^t,$$

donde $\widehat{\Lambda} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p)$ contiene los valores propios ordenados de mayor a menor, y \widehat{U} es la matriz ortogonal de vectores propios asociados.

Repetiendo exactamente el mismo razonamiento que en el caso poblacional, la solución es $\vec{a} = \hat{u}_1$ (primer vector propio de S), y la varianza máxima alcanzada es $\hat{\lambda}_1$. La primera componente principal muestral es por tanto el vector

$$\hat{y}_1 = X_c \hat{u}_1 = HX\hat{u}_1.$$

De manera análoga se obtienen las restantes componentes, imponiendo ortogonalidad (incorrelación muestral) con las anteriores. En general, las p componentes principales muestrales se recogen en la matriz

$$\boxed{\widehat{Y} = X_c \widehat{U} = HX\widehat{U}},$$

donde la columna j de \widehat{Y} es \hat{y}_j , con varianza muestral $\hat{\lambda}_j$ y covarianza muestral nula entre columnas distintas.

11.6. Variabilidad explicada y elección del número de componentes

La variabilidad total muestral se mide por $\text{tr}(S) = \sum_{j=1}^p \hat{\lambda}_j$. La proporción de variabilidad explicada por las q primeras componentes es

$$\frac{\sum_{j=1}^q \hat{\lambda}_j}{\sum_{j=1}^p \hat{\lambda}_j},$$

y se emplean los mismos criterios heurísticos (70 %-80 %, principio de imparcialidad, etc.) descritos para el caso poblacional.

Si se trabaja con la matriz de correlaciones muestral R en lugar de S , las componentes principales normadas muestrales se obtienen de los vectores propios de R . Todos los comentarios sobre la interpretación, calidad de reconstrucción y correlaciones con las variables originales se trasladan sustituyendo Σ por S (o R) y los parámetros poblacionales por sus análogos muestrales.