

EVALUACIÓN DE LA TÉCNICA DESARROLLADA MEDIANTE UN EJEMPLO DE APLICACIÓN

Como aplicación de la metodología usaremos muestras de dos poblaciones de billetes, billetes de Francos Suizos verdaderos y billetes de Francos Suizos falsos, ya utilizados en otros análisis efectuados por Flury y Riedwiel (1983). Las muestras de cada población son de tamaño 100 cada una.

En la siguiente sección se presenta una breve descripción de los datos usados para ilustrar la metodología.

Cabe señalar que por la envergadura del problema, 100 observaciones, de dimensión seis cada una de las dos poblaciones, y dado que a la fecha no se dispone en nuestro medio de un paquete estadístico para resolver un problema de esta naturaleza, se hizo necesario desarrollar un programa computarizado en Lenguaje Fortran.

Descripción de los datos

Se cuenta con las medidas de 6 variables en 200 billetes de banco, 100 de los cuales son verdaderos (proceden de la población uno) y los otros 100 son falsos (proceden de la población dos).

Las variables medidas fueron :

X1 : longitud de los billetes

X2 : Ancho del lado izquierdo de los billetes.

- X3 : Ancho del lado derecho de los billetes.
- X4 : Ancho del margen inferior de los billetes.
- X5 : Ancho del margen superior de los billetes.
- X6 : Longitud de la diagonal medida desde el canto inferior izquierdo hasta el canto superior derecho.

donde:

$$X_i^{(g)} = (X_{i1}^{(g)}, X_{i2}^{(g)}, \dots, X_{i6}^{(g)})$$

$$g = 1, 2 \quad n_1 = n_2 = 100 \quad i = 1, 2, \dots, n_g$$

es el vector que contiene las medidas de las seis variables del i -ésimo billete en el g -ésimo grupo, resultando las matrices de datos muestrales.

Descripción del algoritmo usado en el programa

Para mayor claridad hacemos una descripción resumida de la subrutina FO2AF que se usa dentro del programa computarizado para describir la técnica planteada y que nos permite calcular todos los autovalores y autovectores de la expresión $(B^{-1}A - kI)\vec{b} = 0$ donde A y B son matrices simétricas definidas positivas, usando la reducción de Householder y el algoritmo QL, según Wilkinson, J. y Reinsch, C (1971).

Dado que en general la matriz $B^{-1}A$ no es simétrica, el problema se reduce primero a la obtención de autovalores y autovectores de una matriz simétrica usando el método de descomposición de Cholesky para descomponer la matriz B en matrices triangulares $B = LL'$ donde L es triangular inferior.

$$\text{Como } A\vec{b} = \lambda B\vec{b} \quad (3,1)$$

$$L^{-1}BL^{-1} = I$$

$$AL^{-1}L'\vec{b} = \lambda B\vec{b}$$

$$L^{-1}AL^{-1}L'\vec{b} = \lambda L^{-1}B\vec{b}$$

$$(L^{-1}AL^{-1})(L\bar{b}) = \lambda(L^{-1}B)\bar{b} = \lambda(L\bar{b})$$

$$(L^{-1}AL^{-1})(L\bar{b}) = \lambda(L\bar{b})$$

$$P\bar{d} = \lambda\bar{d} \quad (3,2)$$

Entonces los autovalores de la expresión (3,1) son los de la expresión (3,2) donde $P = L^{-1}BL^{-1}$ y $\bar{d} = L\bar{b}$. Luego, se usa el método de Householder para tridiagonalizar la matriz simétrica P y obtener sus autovalores con el algoritmo QL.

Un autovector \bar{d} de la matriz simétrica P, está relacionado con el autovector \bar{b} de la matriz original B^{-1} mediante la relación:

$$\bar{d} = L\bar{b} \quad (3,3)$$

Puesto que el autovector \bar{d} se obtiene con el algoritmo OL normalizado según $\bar{d}'\bar{d} = 1$, entonces los autovectores del problema original (\bar{b}) se obtienen resolviendo el sistema (3,3) donde \bar{d} es normalizado según $\bar{d}'B\bar{d} = 1$. Osea que el uso de la subrutina FO2AEF permite calcular los autovalores y autovectores de la matriz $S_1^{-1}S_2$ con $A = S_1$ y $B = S_2$ donde S^{-1} es la inversa de la matriz de covarianzas de la muestra del grupo uno.

En lo que sigue, denotamos con r_1, \dots, r_p los autovalores de la matriz $S_1^{-1}S_2$ y con $\bar{C}_1, \dots, \bar{C}_p$ los autovectores asociados a dichos autovalores.

Resultados obtenidos

Recordemos que nuestro problema es docimar la hipótesis de igualdad de las matrices de covarianzas de las dos poblaciones, es decir, $(H_0: \Sigma_1 = \Sigma_2)$, para lo que necesitamos obtener los autovalores de la matriz $S_1^{-1}S_2$.

Consideramos las dos muestras donde los parámetros poblacionales son desconocidos, las medidas descriptivas que son

las estimativas de μ_1, μ_2, Σ_1 y Σ_2 , obtenidas a partir de (Flury, 1983) son las siguientes.

Vector de medias de la muestra uno, billetes verdaderos:

$$\bar{X}^{(1)} = \begin{bmatrix} 214.9690 \\ 129.9430 \\ 129.7200 \\ 8.3050 \\ 10.1710 \\ 141.5150 \end{bmatrix}$$

Vector de medias de la muestra dos, billetes falsos:

$$\bar{X}^{(2)} = \begin{bmatrix} 214.9690 \\ 129.9430 \\ 129.7200 \\ 8.3050 \\ 10.1710 \\ 141.5150 \end{bmatrix}$$

Matriz de covarianzas de la muestra S_1

$$\begin{bmatrix} 0.150372 & 0.058198 & 0.057495 & 0.057130 & 0.014246 & 0.005045 \\ 0.058198 & 0.132603 & 0.0855952 & 0.056648 & 0.048035 & -0.044810 \\ 0.057485 & 0.085952 & 0.126420 & 0.058179 & 0.029673 & -0.023472 \\ 0.057130 & 0.056648 & 0.058179 & 0.413206 & -0.260460 & 0.000628 \\ 0.014246 & 0.048035 & 0.029673 & -0.260460 & 0.415009 & -0.076023 \\ 0.005045 & -0.044810 & -0.023472 & 0.000628 & -0.076023 & 0.200688 \end{bmatrix}$$

Matriz de covarianzas de la muestra S_2

0.124498	0.031602	0.02045	-0.100596	0.019448	0.011619
0.031602	0.064790	0.046783	-0.024041	-0.011918	-0.004987
0.024045	0.046783	0.088727	-0.018574	0.000134	0.034221
-0.100596	-0.024041	-0.024041	1.281313	-0.490192	0.238489
0.019448	-0.011918	0.000134	-0.490192	0.404455	-0.022071
0.011619	-0.004987	0.034221	0.238489	-0.022071	0.311162

Usando la subrutina FO2AEF, cuya descripción fue hecha en la sección anterior, obtenemos los autovalores y autovectores de la matriz $S_1^{-1}S_2$ y que son los siguientes.

$$\begin{aligned}
 r_6 &= 0.283892 & r_5 &= 0.545285 \\
 r_4 &= 0.906826 & r_3 &= 1.052638 \\
 r_2 &= 1.678315 & r_1 &= 6.120406
 \end{aligned}
 \tag{3.5}$$

\bar{C}_6	\bar{C}_5	\bar{C}_4	\bar{C}_3	\bar{C}_2	\bar{C}_1
-0.390511	-1.340457	2.002322	-1.389171	-0.062965	0.977541
-1.194459	3.372635	1.341897	1.030241	0.109882	0.660427
-0.361893	-2.581479	-1.650544	1.910784	1.334965	0.426138
-0.510676	-0.248453	-0.043215	-0.362239	-0.469081	-2.235451
-0.836594	0.023098	-0.746921	-1.320852	0.523342	-1.538391
0.587390	0.626593	0.580347	0.154213	1.934701	-1.059888

Regla de decisión

Como $\hat{\alpha}_1 = r_1 = 6.120406 \gg \hat{\alpha}_6 = r_6 = 0.283892$ se rechaza la hipótesis de igualdad de las matrices de covarianzas, es decir que tenemos el caso de heteroscedastidad entre las matrices de covarianzas de las poblaciones en concurso.

Es necesario señalar que haciendo uso de la docima de homocedasticidad clásica (paquete estadístico Statgraph) se llega

también a rechazar la hipótesis de igualdad de matrices de covarianzas como se puede comprobar en la hoja final del presente capítulo.

Cabe recordar que en el análisis clásico, a partir de la decisión tomada, se está violando un supuesto básico para continuar con futuros trabajos estadísticos tales como discriminación lineal, manova, etc., en tanto que, con la técnica presentada se puede continuar trabajando y por ejemplo obtener las Componentes Principales Generalizadas que no son otra cosa que la generalización de las Componentes Principales en un grupo a Componentes Principales a dos grupos simultáneamente.

También Flury (1983), usando una tabla de un estudio de simulación que aún no ha publicado, afirma que: $P(0.43 \leq r_6 \leq r_1 \leq 2.31) = 0.95$, resultado que confirma la decisión tomada.

Según la decisión tomada, parece razonable continuar con el análisis de los datos, por lo que pasamos a realizar la transformación de Componentes Principales Generalizadas y obtener un nuevo conjunto de variables que estén no correlacionadas en ambos grupos simultáneamente. Para tal fin se definen las combinaciones lineales usando los autovectores de la matriz $S_1^{-1}S_2$. Tales combinaciones lineales son:

$$Y_{\max}^{(g)} = 0.9775X_1^{(g)} + 0.6604X_2^{(g)} + 0.4261X_3^{(g)} - 2.2374X_4^{(g)} - 1.5384X_5^{(g)} - 1.0599X_6^{(g)} \quad (3,6)$$

$$Y_2^{(g)} = 0.629X_1^{(g)} + 0.1098X_2^{(g)} + 1.3349X_3^{(g)} - 0.4691X_4^{(g)} + 0.5233X_5^{(g)} + 1.9347X_6^{(g)}$$

$$Y_3^{(g)} = -1.3891X_1^{(g)} + 1.0302X_2^{(g)} + 1.900781X_3^{(g)} - 0.3622X_4^{(g)} - 1.3808X_5^{(g)} + 0.1542X_6^{(g)}$$

$$Y_4^{(g)} = 2.0023X_1^{(g)} + 1.3418X_2^{(g)} - 1.6505X_3^{(g)} \\ - 0.0432X_4^{(g)} - 0.7469X_5^{(g)} + 0.5803X_6^{(g)}$$

$$Y_5^{(g)} = -1.3405X_1^{(g)} + 3.3726X_2^{(g)} - 2.5815X_3^{(g)} \\ - 0.2484X_4^{(g)} + 0.0231X_5^{(g)} + 0.6265X_6^{(g)}$$

$$Y_{\min}^{(g)} = -0.3905X_1^{(g)} - 1.1944X_2^{(g)} - 0.3619X_3^{(g)} \\ - 0.5107X_4^{(g)} - 0.8366X_5^{(g)} + 0.5874X_6^{(g)}$$

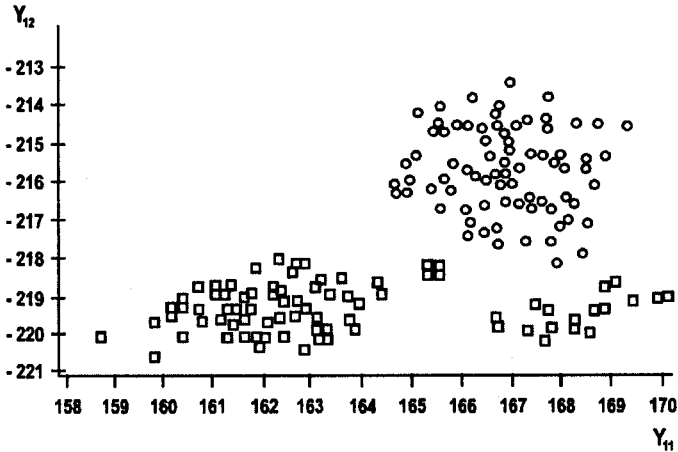
Tienen particular importancia las combinaciones lineales con razones extremas de varianzas (diferentes de uno), es decir las definidas usando los autovectores \vec{C}_1 y \vec{C}_6 asociadas a r_1 y r_6 respectivamente. Usando:

$$\bar{Y}_i^{(g)} = \begin{bmatrix} Y_{i\max}^{(g)} \\ Y_{i\min}^{(g)} \end{bmatrix} \quad \begin{array}{l} i = 1, \dots, n_g \\ g = 1, 2 \\ n_1 = n_2 \end{array} \quad (3,7)$$

Se obtienen las muestras de datos transformados para los billetes falsos y verdaderos respectivamente.

También se observó en el gráfico que para los datos transformados, el grupo de billetes verdaderos tienen forma circular con varianza $C_1^T S_1 C_1 = 1$ mientras que el grupo de billetes falsos tienen la mayor variabilidad en Y_1 y la más pequeña en Y_6 siendo las combinaciones lineales que más distinguen a las poblaciones desde el punto de vista de cociente de varianzas.

Gráfico 1
 Datos transformados de los Billetes de Banco



Círculo: Grupo uno Billetes verdaderos
Cuadrado: Grupo dos Billetes falsos.

Programa usado

Análisis de Componentes Principales Generalizados usando los datos de Flury y Riedwiel (1983)

C CALCULA VECTORES DE MEDIAS, MATRICES DE COVARIANZAS DE LOS DATOS MUESTRALES (MATRICES S_1 Y S_2)

C AUTOVALORES Y AUTOVECTORES DE LA MATRIZ $S_1^{-1}S_2$

C COMBINACIONES LINEALES

C

C ARCHIVO FOR010.DAT CONTIENE LOS DATOS PARA EL GRUPO MUESTRAL DE NOTAS VERDADERAS

C ARCHIVO FOR050.DAT CONTIENE LOS DATOS PARA EL GRUPO MUESTRAL DE NOTAS FALSAS

C

double precision somaX1,somaX2,somaY1,somaY2

double precision X1med,X2med,Y1med,Y2med,somaqX1,
somaqX2

double precision varX1,varX2,a,b,r,v,autove

dimension X1(100,6),X1t(6,100),X2(100,6),X2t(6,100)

dimension somaX1(6),somaX2(6),X1med(6),X2med(6),11(100)

dimension Y1(100,2),Y2(100,2),Y1T(2,100),Y2T(2,100)

dimension somaY1(2),somaY2(2),Y1med(2),Y2med(2)

dimension somaqX1(6,6),somaqX2(6,6)

dimension varX1(6,6),varX2(6,6),A(7,7),B(7,7),R(7),DL(6)

dimension V(7,7),autove(6,2)

C

C Lectura de las matrices de datos

C

do I=1,100

read(10, *) (X1(I,J), J = 1,6)

end do

do I = 1, 100

read(50, *) (X2 (I,J), J = 1,6)

end do

C

C Calculando las transpuestas de las matrices de datos

C

do I = 1, 100

L1(I) = 1

do J = 1, 6

X1 t (J,I) = X1 (I,J)

X2 t (J,I) = X2 (I,J)

end do

end do

C

C Calculando los vectores de medias

```

C
Do I = 1,6
SOMAX1 (I) = 0
X1MED(I) = 0
SOMAX2(I) = 0
X2MED(I)=0
end do
do I = 1,6
Do J = 1,100
SOMAX1(I) = SOMAX1 (I) + X1T(I,J) * L1(J)
SOMAX2(I) = SOMAX2 (I) + X2T(I,J) * L1(J)
END DO
X1MED(I) = SOMAX1(I) / (1.0 * 100)
END DO

```

C

C Calculando las matrices de covarianzas

C

C a)Inicializando las matrices

C

```

do i = 1,6
do j = 1,6
varX1(i,j) = 0.0
varX2(i,j) = 0.0
end do
end do

```

C

C b) Calculando las matrices de suma de cuadrados

C

```

Do I = 1,6
Do J = 1,6
SOMAQX1 (I,J) = 0.0
SOMAQX2 (I,J) = 0.0
end do
end do
do I = 1,6
do j = 1,6

```

```

Do K = 1 , 100
SOMAQX1 (I,J) = SOMAQX1 (I,J) + X1 t(I,K) * X1 (K,J)
SOMAQX2 (I,J) = SOMAQX2 (I,J) + X2 t(I,K) * X2 (K,J)
END DO
SOMAQX1 (j,I) = SOMAQX1 (I,J)
SOMAQX2 (J,I) = SOMAQX2 (I,J)
END DO
END DO
C c) Calculando las matrices de covarianzas
C
DO I = 1 , 6
DO J = I , 6
VARX1 (I,J) = (SOMAQX1 (I,J) - 1.0 * 100 * X1MED(I) *
X1MED(J)) / (1.0 * 99.0)
VARX2 (I,J) = (SOMAQX2 (I,J) - 1.0 * 100 * X2MED(I) *
X2MED(J)) / (1.0 * 99.0)
VARX1 (J,I) = VARX1 (I,J)
VARX2 (J,I) = VARX2 (I,J)
END DO
END DO
C
C d) Redefiniendo las matrices de covarianzas de las dos po-
blaciones para usar la subrutina F02AEF de la biblioteca de NAG
DO 550 I = 1 , 6
DO 550 J = 1 , 6
A(I,J) = VARX2 (I,J)
B(I,J) = VARX1 (I,J)
550 CONTINUE
C
C Calculando los autovalores y autovectores de la matriz
C INV (VARX1) * VARX2 usando la subrutina correspondiente
N=6
TYPE *, B (I,J)
DO I = 1 , N
TYPE *, (B (I,J) , J , N)
END DO

```

```

TYPE *, A (I,J)
DO I = 1 , N
TYPE *, ( A (I,J) , J = 1 , N)
END DO
TYPE *, 'LEU OS DADOS'
IA = 7
IB = 7
IV = 7
IFAIL = 1
CALL F02AEF ( A , IA , B , IB , N , R , V , IV , DL , E , IFAIL )
TYPE *, VOLTOU DA SUBRUTINA
IF ( IFAIL.EQ.0 ) THEN
WRITE ( 33,95 ) ( R (I) , I = 1 , N )
TYPE 95 , ( R (I) , I = 1 , N )
WRITE ( 33,97 ) ( ( V (I,J) , J = 1 , N ) , I = 1 , N )
TYPE 97 , ( ( V (I,J) , J = 1 , N ) , I = 1 , N )
95FORMAT ( 12H0 EIGENVALUES / 1H , 6F18.6 )
97FORMAT ( 13H0 EIGENVECTORS / ( 1H , 6 (2X , F15.6) ) )
ELSE
TYPE 96 , IFAIL
WRITE ( 33 , 96 ) IFAIL
96FORMAT ( 25H0 ERROR In F02AEF IFAIL = I2 )
STOP
END IF
XX = R (6) * R (1)
Do I = 1 , 6
AUTOVE (I,1) = V (I,6)
AUTOVE (I,2) = V (I,1)
END DO

```

C Calculando las combinaciones lineales para obtener las variables de interes usando la técnica de Componentes Principales Generalizados

```

C Do I = 1 , 100
Do J = 1 , 2

```

Y1 (I,J) = 0.0

Y2 (I,J) = 0.0

END DO

END DO

C

DO I = 1 , 100

DO J = 1 , 2

DO K=1 , 6

Y1 (I,J) = Y1 (I,J) + X1 (I,K) * AUTOVE (K,J)

Y2 (I,J) = Y2 (I,J) + X2 (I,K) * AUTOVE (K,J)

END DO

END DO

END DO

C

C Calculando las matrices transpuestas de las variables transformadas

do I = 1 , 100

do J = 1 , 2

Y1 T (J,I) = Y1 (I,J)

Y2 T (J,I) = Y2 (I,J)

END DO

END DO

C

C Calculamos el vector de medias de las variables transformadas (Y)

C

do I = 1 , 2

somaY1 (I) = 0.0

Y1med(I) = 0.0

somaY2(I) = 0.0

Y2med(I) = 0.0

do J = 1 , 100

somaY1 (I) = somaY1 (I) + Y1 (I,J) * L1 (J)

SomaY2 (I) = somaY2 (I) + Y2 T (I,J) * L1 (J)

end do

```

Y1med (i) = somaY1 (I) / (1.0 * 100 )
Y2med (i) = somaY2 (I) / (1.0 * 100 )
end do
C Imprimiendo y generando algunos resultados
C
Type 10
Write ( 33 , 10 )
10Format ( / , 15X , 'matriz de covarianzas da amostra um' )
do I = 1 , 6
type 15 , ( varX1 (i,j) , j = 1 , 6 )
write ( 33 , 15 ) ( varX1 (i,j) , j = 1 , 6 )
15Format ( / 15 X , 6 ( f 10.6 , 3X ) )
end do
type 20
write ( 33 , 20 )
20Format ( / , 15X , 'matriz de covarianzas da amostra dois' )
do I = 1 , 6
type 25 , ( varX2 (i,j) , j = 1 , 6 )
write ( 33 , 25 ) ( varX2 (i,j) , j = 1 , 6 )
25Format ( / , 15 X , 6 ( f 10.6 , 3X ) )
end do
type 31
write ( 33 , 31 )
31Format ( / 15X , matriz dados transformados das amostras
um e doi' )
do I = 1 , 100
type 35 , ( Y1 (i,j) , j = 1 , 2 ) , ( Y2 (i,j) , j = 1 , 2 )
write ( 34 , 35 ) ( Y1 (i,j) , j = 1 , 2 ) , ( Y2 (i,j) , j = 1 , 2 )
35Format ( / , 10X , 100 ( f 15.6 , 2X ) )
end do
type 30
write ( 33 , 30 )
30Format ( / , 15X , 'vectores de medias das variaveis trans-
formada' )
do I = 1 , 2
type 32 , Ymed (I) , Y2med(I)

```

```
write (33 , 32 ) Y1med(I) , Y2med(I)
31Format ( / , 2 (3X , f 15.6 ) )
END DO
STOP
END
```