TRABAJOS DE ESTADÍSTICA Y
DE INVESTIGACIÓN OPERATIVA
Vol. 35. Núm. 1, 1984, pp. 65 a 82

CRITERIO PARA DETECTAR OUTLIERS EN POBLACIONES NORMALES BIVARIANTES

Joaquín Muñoz García
Departamento de Estadística y de Investigación Operativa.
Universidad de Sevilla.

Resumen

Damos un procedimiento de detección de outliers para muestras procedentes de poblaciones normales bivariantes, que viene dado por el cuadrado de la distancia entre matrices de sumas de cuadrados y sumas de productos de observaciones muestrales, la cual se ha obtenido a partir de la forma métrica diferencial de MAAS.

1. INTRODUCCIÓN

Para definir el cuadrado de la distancia entre matrices de sumas de cuadrados y sumas de productos nos basamos en la distancia geodésica entre matrices simétricas definidas positivas, utilizando para ello la forma métrica diferencial dada por MAAS (1955).

1.1 Definición. Sea K un espacio conexo sobre el que se ha definido una forma métrica diferencia $(ds)^2$; la distancia entre dos puntos $X, Y \in K$ viene dada por.

$$d(X, Y) = \inf \int_X^Y ds$$

(*) Recibido, Julio, 1981

1.1 Teorema. La función $d(X, Y)$ verifica las siguientes propiedades

$$d(X, Y) \geqslant 0$$
$$d(X, Y) = 0 \Leftrightarrow X = Y$$
$$d(X, Y) = d(Y, X)$$
$$d(X, Z) \leqslant d(X, Y) + d(Y, Z)$$

y la topologia definida por d es equivalente a la topología inicial del espacio K. [KOBAYASHI, NOMIZU (1963)]

1.2 Teorema. Las matrices simétricas definidas positivas de dimensión $p \times p$ forman un cono convexo en el espacio euclideo $R^{p(p+1)/2}$.

DEMOSTRACIÓN. Basta considerar, cada matriz simétrica como un punto del espacio euclideo $p(p + 1)/2$ – dimensional.

1.3 Teorema. Sean A y B dos matrices simétricas definidas positivas. Tomando como forma diferencial la dada por Maas, la distancia geodésica entre estas matrices viene dada por

$$d(A, B) = \left(\sum_{i=1}^{p} (\log \lambda_i)^2 \right)^{1/2}$$

donde λ_i son las raíces de la ecuación determinante $|B - \lambda A| = 0$.

La demostración de este teorema, se puede ver en MUÑOZ (1980).

2. DISTRIBUCIÓN DEL ESTADÍSTICO DISTANCIA ENTRE MATRICES DE SUMAS DE CUADRADOS Y SUMAS DE PRODUCTOS

Aplicaremos los resultados anteriores a las matrices de sumas de cuadrados y sumas de productos de observaciones muestrales, pero veamos en primer lugar que esto es posible.

Sea

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1n} \\ x_{21} & x_{22} & \cdots\cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots\cdots & x_{pn} \end{pmatrix}$$

una muestra aleatoria simple procedente de una población $N_p(\mu, \Sigma)$ no singular, entonces la matriz de s. c. y s. p. de observaciones muestrales para esta muestra de tamaño n viene dada por

$$S_{(n)} = X(I - \tfrac{1}{n} E_{nn})X'$$

que es una matriz simétrica y además si no conocemos los parametros poblacionales, $S_{(n)}$ se distribuye según una ley de Wishart p-dimensional con $n - 1$ grados de libertad y matriz asociada $\Sigma[W_p(n - 1, \Sigma)]$ ANDERSON (1958) y si se tiene que $n > p$, entonces $S_{(n)}$ es definida positiva con probabilidad uno, lo cual se puede representar simbólicamente como

$$\int_{S_{(n)} > 0} W_p(n - 1, \Sigma)\, dS_{(n)} = 1$$

Y con esto podemos afirmar que las matrices de s. c. y s. p. de observaciones muestrales cumplen las condiciones exigidas para poder definir la función distancia $d(S_{(n-k)}, S_{(n)})$, con ($n > p$ y $n - k > p$) salvo conjuntos de medida nula, en la forma.

$$d(S_{(n-k)}, S_{(n)}) = \left(\sum_{i=1}^{p} (\log \lambda_i)^2 \right)^{1/2} \qquad (1)$$

donde $S_{(n-k)}$ es la matriz de s. c. y s. p. de $n - k$ observaciones y los λ_i son las soluciones de la ecuación determinante

$$|S_{(n-k)} - \lambda S_{(n)}| = 0$$

Como estamos trabajando con matrices de tipo aleatorio, la distancia definida en (1), podemos considerarla como una variable aleatoria y por tanto podemos hablar de la distribución de la variable aleatoria distancia. Para calcular dicha distribución vamos a basarnos en los siguientes lemas y teoremas.

2.1 Lema. Sea X una muestra aleatoria simple de tamaño $n(n > p)$ procedente de una población $N_p(\mu, \Sigma)$ no singular. La matriz de s. c. y s. p. de la muestra $X, S_{(n)}$ puede descomponerse de la siguiente forma.

$$S_{(n)} = S_{(n-k)} + S_{(h)} + \frac{k(n-k)}{n} (\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})$$

donde $S_{(n-k)}$ es la matriz de s. c. y s. p. de $n - k(n - k > p)$ observaciones y $\bar{X}_{(n-k)}$ su vector media, siendo $S_{(k)}$ la matriz de s. c. y s. p. de las $k(k > p)$ observaciones restantes y $\bar{X}_{(k)}$ su vector media.

Además

$$S_{(k)} + \frac{k(n-k)}{n}(\bar{X}_{(n-k)} - \bar{X}_{(h)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})' \in W_p(K, \Sigma)$$

Demostración.

$$S_{(n)} = X(I - \tfrac{1}{n}E_{nn})X' = (X_{(n-k)}|X_{(k)})(I - \tfrac{1}{n}E_{nn})(X_{(n-k)}|X_{(k)})'$$
$$\begin{aligned} S_{(n)} = X(I - \tfrac{1}{n}E_{nn})X' &= X_{(n-k)}X'_{(n-k)} + X_{(k)}X'_{(h)} - \\ &- \tfrac{1}{n}[X_{(n-k)}E_{(n-k),(n-k)}X'_{(n-k)} + X_{(n-k)}E_{(n-k),k}X'_{(k)} + \\ &+ X_{(k)}E_{k,(n-k)}X'_{(n-k)} + X_{(k)}E_{kk}X'_{(k)} \end{aligned}$$

Sumando y restando

$$\frac{1}{n-k}X_{(n-k)}E_{(n-k),(n-k)}X'_{(n-k)}; \qquad \frac{1}{k}X_{(x)}E_{k,k}X'_{(k)}$$

se obtiene

$$S_{(n)} = S_{(n-k)} + S_{(k)} + \frac{k(n-k)}{n}(\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})'$$

Por último como

$$S_{(x)} \in W_p(K, \Sigma)$$

y

$$\frac{k(n-k)}{n}(\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})'$$

se distribuye según una ley pseudo-Wishart p-dimensional con un grado de libertad y matriz asociada Σ. Y al ser ambas variables independientes [WILKS (1962)] se tendrá que

$$S_1 = S_{(k)} + \frac{k(n-k)}{n}(\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})'$$

se distribuye según una ley Wishart p-dimensional con k grados de libertad y matriz asociada Σ.

2.2 Lema. Sea S_1 una matriz $p \times p$ que se distribuye según una ley $W_p(k, \Sigma)$ y $S_{(n-k)}$ una matriz $p \times p$ que se distribuye según una ley $W_p(n - k - 1, \Sigma)$ independiente de la anterior y sea $S_{(n)} = S_{(n-k)} + S_1$. Entonces la distribución conjunta de las raíces de la ecuación determinante

$$|S_{(n-l)} - \lambda S_{(n)}|0$$

es de la forma

$$f(\lambda_1, \lambda_2, \ldots, \lambda_p) = A \prod_{i=1}^{p} \prod_{j=i+1}^{p} (\lambda_j - \lambda_i) [\prod_{i=1}^{p} \lambda_i]^{1/2(n-k-p-2)} \times$$

$$\times \left[\prod_{i=1}^{p} (1 - \lambda_i) \right]^{1/2(k-p-1)}$$

donde

$$0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_p \leqslant 1$$

y

$$A = \pi^{p/2} \prod_{i=1}^{p} \frac{\Gamma[\frac{1}{2}(n-i)]}{\Gamma[\frac{1}{2}(n-k-i)]\Gamma[\frac{1}{2}(k-i+1)]\Gamma[\frac{1}{2}(p-i+1)]}$$

Es un caso particular de los resultados debidos a Hsu (1939), Fisher (1939), Roy(1939), y Khrisnaiah (1978).

2.1 Teorema. Sea X una muestra aleatoria simple de tamaño n ($n > 2$) extraida de una población con distribución $N_2(\mu, \Sigma)$ no singular. Sea $S_{(n)}$ la matriz de s. c. y s. p. de las observaciones muestrales y $S_{(n-k)}$ la matriz de s. c. y s. p. de $(n-k)$ observaciones ($k > 2$).

La función de densidad del estadistico cuadrado de la distancia

$$U = d^2(S_{(n-k)}, S_{(n)}) = (\log \lambda_1)^2 + (\log \lambda_2)^2$$

donde λ_1 y λ_2 son las soluciones de la ecuación determinante

$$|S_{(n-k)} - \lambda S_{(n)}| = 0$$

viene dada por

$$f(u) = \begin{cases} A e^{-\sqrt{e}\,\frac{n-k-2}{2}} \int_{\sqrt{2}-1}^{1} \left(e^{-\sqrt{u}\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\frac{2w}{1+w^2}}\right) e^{-\sqrt{u}(n-k-2)\frac{w}{1-w}} \,{}_1\mp \\ \left[\left(1 - e^{-\sqrt{u}\frac{2w}{1+w^2}}\right)\left(1 - e^{-\sqrt{u}\frac{1-w^2}{1+w^2}}\right)\right]^{\frac{k-3}{2}} dw \quad 0 < u < \infty \\ 0 \text{ en el resto} \end{cases}$$

siendo

$$A = \frac{\sqrt{\pi}\,\Gamma[\frac{1}{2}(n-1)]\,\Gamma[\frac{1}{2}(n-2)]}{\Gamma(\frac{1}{2}[n-k-1)]\,\Gamma[\frac{1}{2}(n-k-2)]\,\Gamma(\frac{k}{2})\,\Gamma(\frac{k-1}{2})}$$

Demostración. Partiendo de la función de densidad dada en el Lema 2.2 para el caso en que $p = 2$.

Y realizando las trasformaciones

$$\begin{aligned} X &= -\ln \lambda_1 \\ Y &= -\ln \lambda_2 \end{aligned} \qquad 0 < Y < X < \infty$$

y a continuación

$$\begin{aligned} \theta &= X^2 \\ \varphi &= Y^2 \end{aligned}$$

se obtiene

$$f(\theta, \varphi) = \tfrac{1}{4}(e^{-\sqrt{\varphi}} - e^{-\sqrt{\theta}})\,e^{-(\sqrt{\theta}+\sqrt{\varphi})\frac{1}{2}(n-k-2)} \times$$

$$\times\,[(1 - e^{-\sqrt{\varphi}})(1 - e^{-\sqrt{\theta}})]^{\frac{k-3}{2}}(\theta\varphi)^{-1/2}$$

$$0 < \varphi < \theta < \infty$$

seguidamente realizamos los cambios de variables

$$\begin{aligned} U &= \theta + \varphi \\ t &= \theta \end{aligned} \qquad 0 < u < \infty \qquad \tfrac{u}{2} < t < u$$

$$Z = \sqrt{\frac{2(U-t)}{U}} \qquad 0 < Z < 1$$

y por último la trasformación

$$\sqrt{1 - \frac{z^2}{2}} = \left(1 + \frac{Z}{\sqrt{2}}\right) w$$

dándonos:

$$f(u) = A e^{-\sqrt{u}\,\frac{n-k-2}{2}+1} \Big|_{\cdot 2-1} (e^{-\sqrt{u}\,\frac{1}{1+w^2}} - e^{-\sqrt{u}\,\frac{1}{1+w^2}})\,e^{-\sqrt{u}(n-k-2)w\,\frac{1}{1+w^2}} \times$$

$$\times [(1 - e^{-\sqrt{u}\frac{2w}{1+w^2}})(1 - e^{-\sqrt{u}\frac{1-w^2}{1+w^2}})]^{\frac{k-1}{2}} \frac{dw}{1 + w^2} \quad 0 < u < \infty$$

siendo

$$A = \frac{\sqrt{\pi}\,\Gamma[\frac{1}{2}(n-1)]\,\Gamma[\frac{1}{2}(n-2)]}{\Gamma(\frac{1}{2}[n-k-1)]\,\Gamma[\frac{1}{2}(n-k-2)]\,\Gamma(\frac{k}{2})\,\Gamma(\frac{k-1}{2})}$$

3. CONSTRUCCIÓN DE LA FUNCIÓN g

Dada la dificultad de tabulación de la función de distribución F debido a la función de densidad f vamos a construir una función de densidad g cuya función de distribución G será dominada estocásticamente por F siendo más fácil su tabulación. Estos resultados son consecuencia de los siguientes teoremas y programas.

3.1 Teorema. La función

$$f(u) = Ae^{-\sqrt{u}\frac{n-k-2}{2}} \int_{\sqrt{2}-1}^{1} (e^{-\sqrt{u}\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\frac{2w}{1+w^2}})e^{-\sqrt{u}(n-k-2)\frac{w}{1+w^2}} \times$$

$$\times \frac{1}{1+w^2}\left[(1 - e^{-\sqrt{u}\frac{2w}{1+w^2}})(1 - e^{-\sqrt{u}\frac{1-w^2}{1+w^2}})\right]^{\frac{k-3}{2}} dw \quad 0 < u < \infty$$

siendo

$$A = \frac{\sqrt{\pi}\,\Gamma[\frac{1}{2}(n-1)]\,\Gamma[\frac{1}{2}(n-2)]}{\Gamma(\frac{1}{2}[n-k-1)]\,\Gamma[\frac{1}{2}(n-k-2)]\,\Gamma(\frac{k}{2})\,\Gamma(\frac{k-1}{2})}$$

está acotada superiormente por la función

$$g_1(U) = Ae^{-\sqrt{u}\frac{n-k-2}{2}}\left(1 - e^{-\sqrt{u}}\right)^{k-2} \qquad 0 < U < \infty$$

Este resultado se obtiene aplicando sucesivamente la siguiente propiedad de la integración.

Sean $h(x)$ y $t(x)$ integrales sobre el intervalo (a, b), sea M el supremo de $h(x)$ en (a, b). Sea, además $t(x) \geqslant 0$ en (a, b). En estas condiciones se tiene:

$$\int_a^b h(x)t(x)dx \leqslant M\int_a^b t(x)dx$$

Y a partir de esta función g_1 construimos la función de densidad dada por el siguiente teorema.

3.2 Teorema. La función $g(x)$ definida por

$$g(x) = \frac{1}{2\sum_{j=0}^{k-2}\left[\binom{k-2}{j}(-1)^j \frac{1}{\left(\frac{n-k-2}{2}+j\right)^2}\right]} e^{-\sqrt{x}\left(\frac{n-k-2}{2}\right)}\left(1-e^{-\sqrt{x}}\right)^{k-2} \qquad 0 < x < \infty$$

con $k > 2$ y $n - k > 2$ es función de densidad.

A continuación damos dos programas de ordenador. El primero nos sirve para la tabulación de la función $G(x)$, estando realizado en la situación $n = 15$ y $k = 5$. Y el segundo incluye las representaciones de $f(u)$ y $g(x)$ en el caso particular $n = 55$ y $k = 10$.

Tabulación de la función de distribución $G(x)$:

```
      DOUBLE PRECISION FUNCTION SIN3NI(FX.A.E.REL.MAXIT.FX.◄)          SIM3NI
      IMPLICIT DOUBLE PRECISION (A-H, O-Z)
    C------------------------------------------------------------------SIM3NI
    C INTEGRACION NUMERICA UTILIZANDO LA REGLA 3/8 DE SIMPSON            SIM3NI
    C------------------------------------------------------------------SIM3NI
    C DIMENSION A(2)                                                    SIM3NI
      LOGICAL REL                                                       SIM3NI
      PREV=0                                                            SIM3NI
    C------------------------------------------------------------------SIM3NI
    C INICIALIZAR H.X.N.M.S                                             SIM3NI
    C------------------------------------------------------------------SIM3NI
      H=(A(2)-A(1))/3.                                                  SIM3NI
      X=A(1)                                                            SIM3NI
      N=0                                                               SIM3NI
      M=3                                                               SIM3NI
      S=0.                                                              SIM3NI
    C------------------------------------------------------------------SIM3NI
    C CICLO HASTA EL MAXIMO NUMERO DE EVALUACIONES                      SIM3NI
    C------------------------------------------------------------------SIM3NI
      DO 3 J=1,MAXIT                                                    SIM3NI
    C------------------------------------------------------------------SIM3NI
    C CICLO PARA EL NUMERO DE EVALUACIONES                              SIM3NI
    C------------------------------------------------------------------SIM3NI
      DO 1 I=N.N                                                        SIM3NI
      R=3.                                                              SIM3NI
    C------------------------------------------------------------------SIM3NI
    C DETERMINACION DEL COEFICIENTE R                                   SIM3NI
    C------------------------------------------------------------------SIM3NI
      IF(MOD(I 3).EQ.2*N) R=N+1.                                        SIM3NI
    C------------------------------------------------------------------SIM3NI
    C SUMA DE LAS EVALUACIONES DE LA FUNCION                            SIM3NI
    C------------------------------------------------------------------SIM3NI
```

```
      S=S+FX(X,FK)*R                                                    SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C INCREMENTAR X                                                     SIM3NI
    C------------------------------------------------------------------ SIM3NI
      1 X=X+H                                                           SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C OBTENCION DEL NUEVO VALOR DE INTEGRACION                          SIM3NI
    C------------------------------------------------------------------ SIM3NI
      SIM 3NI=S*N*.375D0/(N+1.DO)                                       SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C TEST PARA EL PRIMER CICLO                                         SIM3NI
    C------------------------------------------------------------------ SIM3NI
      IF(N.EQ.0) GO TO 2                                                SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C PARA LOS CICLOS DISTINTOS DEL PRIMERO,
      MITAD DE H Y DOBLE DE M                                           SIM3NI
    C------------------------------------------------------------------ SIM3NI
      H=H*.5DO                                                          SIM3NI
      M=2*M                                                             SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C REVISION DEL ERROR DE CONTROL                                     SIM3NI
    C------------------------------------------------------------------ SIM3NI
      R=SIM3NI-PREV                                                     SIM3NI
      IF(REL) R=R/SIM3NI                                                SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C SI EL ERROR ESTA DENTRO DE LOS LIMITES DADOS,FIN                  SIM3NI
    C------------------------------------------------------------------ SIM3NI
      IF(DABS(R).LT.E) GO TO 4                                          SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C FIJAR UN NUEVO VALOR DE INTEGRACION                               SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C 2 PREV=SIM3NI                                                     SIM3NI
    C   N=1                                                             SIM3NI
    C------------------------------------------------------------------ SIM3NI
    C OBTENER EL NUEVO LIMITE INFERIOR PARA LA EVALUACION DE LA FUNCION SIM3NI
    C------------------------------------------------------------------ SIM3NI
      3 X X=A(1)+.5DO*H                                                 SIM3NI
        RETURN 7                                                        SIM3NI
      4 RETURN                                                          SIM3NI
        END                                                             SIM3NI

 END OF COMPILATION:      NO DIAGNOSTICS.


        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        EXTERNAL F1
        LOGICAL REL
        COMMON N,K
        N=15
        K=5
        DIMENSION A(2)
        B=4./(N-K-2.)**2
        K2=K-2
        DO 1 IR=1,K2
```

```
      1 B=B+(-1)**IR*(FACT(K2)/(FACT(IR)*FACT(K2-IR)))*
        *(4./(N-K-2+2*IR)**2)
        A(1)=0
        A(2)=9
        E=1.D-6
        MAXIT=200
        REL=.TRUE.
        DO 3 I=1,200
        A(2)=A(2)+1.D-1
        AINT=SIM3NI(F1,A,E,REL,MAXIT,1.D,$2)
        F=AINT/(2*B)
        PRINT 6,A(2),F
      6 FORMAT(10X,'F('.D9.4,')=',D20.10)
        GO TO 3
      2 PRINT 4,N,K
      4 FORMAT('    ERROR EN LA INTEGRAL PARA N=',I4,' K='.14)
      3 CONTINUE
        STOP
        END

        END OF COMPILATION:       NO DIAGNOSTICS.

        DOUBLE PRECISION FUNCTION FACT(M)
        COMMON N,K
        IF(M) 1,3,4
      1 PRINT 2,N,K
      2 FORMAT(' ARGUMENTO NEGATIVO PARA UN FACTORIAL .N='.I4,' K= ',I4)
        STOP
      3 FACT=1
        RETURN
      4 FACT=1
        DO 5 I=1,M
      5 FACT=FACT*1
        RETURN
        END

END OF COMPILATION:       NO DIAGNOSTICS.

        DOUBLE PRECISION FUNCTION F1(U,FK)
        IMPLICIT DOUBLE PRECISION (A-H,0-Z)
        COMMON N,K
        F1=(DEXP(-DSQRT(U)*((N-K-2.)/2.)))*(1-DEXP(-DSQRT(U)))**(K-2)
        RETURN
        END

END OF COMPILATION:       NO DIAGNOSTICS.

        DOUBLE PRECISION FUNCTION SIM3NI(FX,A,E,REL,MAXIT,FK,$)
```

```
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      DIMENSION A(2)
      LOGICAL REL
      PREV=0.
      H=(A(2)-A(1))/3.
      X=A(1)
      N=0
      M=3
      S=0
      DO 3 J=1,MAXIT
      DO 1 I=N,M
      R=3.
      IF(MOD(I,3).EQ.2*N) R=N+1.
      2=S+FX(X,FK)*R
    1 X=X+H
      SIM 3NI=S*H*.375DO/(N+1.DO)
      IF(N.EQ.0) GO TO 2
      H=H*.5DO
      M=2*M
      R=SIM 3NI-PREV
      IF(REL) R=R/SIM3NI
      IF(DABS(R).LT.E) GO TO 4
    2 PREV=SIM3NI
      N=1
    3 X=A(1)+.5D0*H
      RETURN 7
    4 RETURN
      END
```

END OF COMPILATION: NO DIAGNOSTICS.

```
      DOUBLE PRECISION FUNCTION FACT(M)
      COMMON N,K
      IF(M) 1,3,4
    1 PRINT 2,N,K
    2 FORMAT(' ARGUMENTO NEGATIVO PARA UN FACTORIAL ,N='.I4,' K=',I4)
      STOP
    3 FACT=1
      RETURN
    4 FACT=1
      DO 5 I=1,M
    5 FACT=FACT*I
      RETURN
      END
```

END OF COMPILATION: NO DIAGNOSTICS.

```
      DOUBLE PRECISION FUNCTION F1(T,FK)
      IMPLICIT DOUBLE PRECISON (A-H,O-U,W-Z)
```

```
      COMMON N,K,U
      R=DEXP (-DSQRT(U)*(1. -T**2)/(1. +T**2))
      S=DEXP(-DSQRT(U)*2. *T/(1. +T**2))
      Q=DEXP(-DSQRT(U)*(N-K-2)*T*(1. -T)/(1. +T**2))
      F1=(R-S)*Q*((1. -S)*(1. -R))**((K-3)/2. )/(1. +T**2)
      RETURN
      END
```

```
END OF COMPILATION:        NO DIAGNOSTICS.
```

Representación gráfica de las funciones $f(u)$ y $g(x)$.

```
C ---------------------------------------------------------------------- SIM3NI
C COMPARACION Y REPRESENTACION DE LAS FUNCIONES DE DENSIDAD DE LA VARIABLE
C DISTANCIA AL CUADRADO F(U) Y LA DADA EN EL TEOREMA ,3.2
C (EN EL PROGRAMA, F(U) ES F, Y LA DEL TEOREMA ES G)
C ---------------------------------------------------------------------- SIM3NI
      IMPLICIT DOUBLE PRECISION(A-H,O-U,W-Z)                              SIM3NI
      COMMON N,K,U                                                        SIM3NI
      DIMENSION A(2),V(61,81),U1(61),F2(61),G1(61)                        SIM3NI
      EXTERNAL F1                                                         SIM3NI
      LOGICAL REL                                                         SIM3NI
      N=55                                                                SIM3NI
      U=0                                                                 SIM3NI
      K=10                                                                SIM3NI
      DO 10 I=1,61                                                        SIM3NI
      DO 10 J=1,81                                                        SIM3NI
   10 V(I,J)=' '                                                          SIM3NI
      DO 11 I=1,61                                                        SIM3NI
   11 V(I,1)='I'                                                          SIM3NI
      DO 12 I=1,81                                                        SIM3NI
   12 V(1,I)='-'                                                          SIM3NI
      G1(1)=0                                                             SIM3NI
      F2(1)=0                                                             SIM3NI
      U1(1)=0                                                             SIM3NI
C ---------------------------------------------------------------------- SIM3NI
C CALCULO DE LAS CONSTANTES B Y C                                         SIM3NI
C ---------------------------------------------------------------------- SIM3NI
      B=4./(N-K-2.)**2                                                    SIM3NI
      K2=K-2                                                              SIM3NI
      DO 1 IR=1,K2                                                        SIM3NI
    1 B=B+(-1)**IR*(FACT(K2)/(FACT(IR)*FACT(K2-IR)))*                     SIM3NI
     *(4./(N-K-2+2*IR)**2)                                                SIM3NI
      C=FACT(N-3)/(4*FACT(K2)*FACT(N-K-3)                                 SIM3NI
C ---------------------------------------------------------------------- SIM3NI
C CALCULO DE LAS FUNCIONES F Y G                                          SIM3NI
C ---------------------------------------------------------------------- SIM3NI
      A(1)=DSQRT(2.D)-1.D                                                 SIM3NI
      A(2)=1.D                                                            SIM3NI
      E=1.D-6                                                             SIM3NI
```

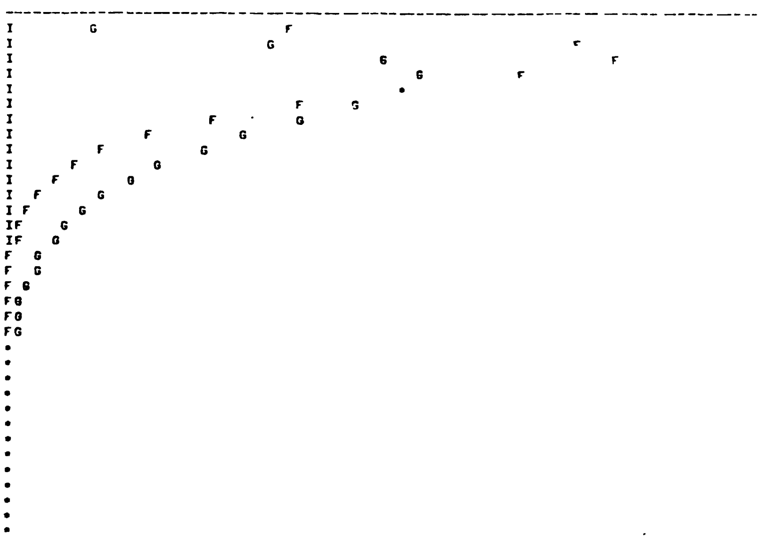
```
      MAXIT=200                                                         SIM3NI
      REL=.TRUE.                                                        SIM3NI
      PRINT 9                                                           SIM3NI
    9 FORMAT(1X,/,1X, 'VALORES COMPARATIVOS DE LAS FUNCIONES F Y G',/)  SIM3NI
      DO 2 I=1,60                                                       SIM3NI
      U=U+.25/10                                                        SIM3NI
      AINT=SIM3NI(F1,A,E,REL,MAXIT,1,D,$3)                              SIM3NI
      F=C*AINT*DEXP(-DSQRT(U)*(N-K-2)/2. )                              SIM3NI
      G=((DEXP(-DSQRT(U)*((N-K-2)/2.)))*(1.-DEXP(-DSQRT(U)))--(K-2))    SIM3NI
      G=G/(2*8)                                                         SIM3NI
      II=I+1                                                            SIM3NI
      JF=(90*F)/10+1                                                    SIM3NI
      JG=(90*G)/10+1                                                    SIM3NI
      IF(JF.NE.JG) GO TO 22                                             SIM3NI
      V(II,JF)='*'                                                      SIM3NI
      GOTO 23                                                           SIM3NI
   22 V(II,JF)='F'                                                      SIM3NI
      V(II,JG)='G'                                                      SIM3NI
   23 G1(II)=G                                                          SIM3NI
      F2(II)=F                                                          SIM3NI
      U1(II)=U                                                          SIM3NI
      PRINT 6,U,F,G                                                     SIM3NI
    6 FORMAT(1X,'U= 'D9.3,10X,'10X,'F(U)= 'D12.67,10X,'G(X)= 'D12.6,/)  SIM3NI
      GOTO 2                                                            SIM3NI
    3 PRINT 7,U                                                         SIM3NI
    7 FORMAT(1X,'ERROR EN U=' D9.3)                                     SIM3NI
    2 CONTINUE                                                          SIM3NI
C ------------------------------------------------------------------------SIM3NI
C GRAFICA DE LAS FUNCIONES DE DENSIDAD F Y G                            SIM3NI
C ------------------------------------------------------------------------SIM3NI
      PRINT 31                                                          SIM3NI
   31 FORMAT(1H1)                                                       SIM3NI
      PRINT 30,((U1(I),F2(I),G1(I),(V(I,J),J=1,81)),I=1,61)             SIM3NI
   30 FORMAT(1X,D9.3,1X,D12.5,1X,D12.5,10X,81A1)                        SIM3NI
      STOP                                                              SIM3NI
      END                                                               SIM3NI
```

```
END OF COMPILATION:      NO DIAGNOSTIC.
```

```
U= .250-001    F(U)= .336041+001    G(X)= .105735+001
U= .500-001    F(U)= .686483+001    G(X)= .321111+001
U= .750-001    F(U)= .722893+001    G(X)= .455085+001
U= .100+000    F(U)= .612100+001    G(X)= .492349+001
U= .125+000    F(U)= .470783+001    G(X)= .467910+001
U= .150+000    F(U)= .345121+001    G(X)= .413799+001
U= .175+000    F(U)= .246709+001    G(X)= .350369+001
U= .200+000    F(U)= .174066+001    G(X)= .288556+001
U= .225+000    F(U)= .122053+001    G(X)= .233355+001
U= .250+000    F(U)= .854046+000    G(X)= .186422+001
U= .275+000    F(U)= .597888+000    G(X)= .147708+001
```

U= .300+000	F(U)= .419435+000	G(X)= .116392+001
U= .325+000	F(U)= .295162+000	G(X)= .913881+000
U= .350+000	F(U)= .208492+000	G(X)= .715980+000
U= .375+000	F(U)= .147886+000	G(X)= .560267+000
U= .400+000	F(U)= .105359+000	G(X)= .438223+000
U= .425+000	F(U)= .754015-001	G(X)= .342802+000
U= .450+000	F(U)= .542088-001	G(X)= .268300+000
U= .475+000	F(U)= .391504-001	G(X)= .210167+000
U= .500+000	F(U)= .284028-001	G(X)= .164808+000
U= .525+000	F(U)= .206973-001	G(X)= .129402+000
U= .550+000	F(U)= .151480-001	G(X)= .101744+000
U= .575+000	F(U)= .111337-001	G(X)= .801179-001
U= .600+000	F(U)= .821729-002	G(X)= .631876-001
U= .625+000	F(U)= .608932-002	G(X)= .499159-001
U= .650+000	F(U)= .453015-002	G(X)= .394973-001
U= .675+000	F(U)= .338309-002	G(X)= .313059-001
U= .700+000	F(U)= .253585-002	G(X)= .248553-001
U= .725+000	F(U)= .190763-002	G(X)= .197675-001
U= .750+000	F(U)= .144006-002	G(X)= .157477-001
U= .775+000	F(U)= .109078-002	G(X)= .125666-001
U= .800+000	F(U)= .828939-003	G(X)= .100448-001
U= .825+000	F(U)= .631965-003	G(X)= .804248-002
U= .850+000	F(U)= .483291-003	G(X)= .644984-002
U= .875+000	F(U)= .370707-003	G(X)= .518098-002
U= .900+000	F(U)= .285181-003	G(X)= .416840-002
U= .925+000	F(U)= .220010-003	G(X)= .335901-002
U= .950+000	F(U)= .170201-003	G(X)= .271099-002
U= .975+000	F(U)= .132022-003	G(X)= .219134-002
U= .100+001	F(U)= .102674-003	G(X)= .177396-002
U= .102+001	F(U)= .800521-004	G(X)= .143821-002
U= .105+001	F(U)= .625686-004	G(X)= .116770-002
U= .107+001	F(U)= .490209-004	G(X)= .949436-003
U= .110+001	F(U)= .384964-004	G(X)= .773054-003
U= .112+001	F(U)= .303003-004	G(X)= .630312-003
U= .115+001	F(U)= .239020-004	G(X)= .514626-003
U= .117+001	F(U)= .188955-004	G(X)= .420735-003
U= .120+001	F(U)= .149691-004	G(X)= .344427-003
U= .122+001	F(U)= .118830-004	G(X)= .282323-003
U= .125+001	F(U)= .945195-005	G(X)= .231711-003
U= .127+001	F(U)= .753296-005	G(X)= .190409-003
U= .130+001	F(U)= .601502-005	G(X)= .156662-003
U= .132+001	F(U)= .481189-005	G(X)= .129051-003
U= .135+001	F(U)= .383641-005	G(X)= .106432-003
U= .137+001	F(U)= .309614-005	G(X)= .878806-004
U= .140+001	F(U)= .249006-005	G(X)= .726456-004
U= .142+001	F(U)= .200602-005	G(X)= .601195-004
U= .145+001	F(U)= .161875-005	G(X)= .498084-004
U= .147+001	F(U)= .130837-005	G(X)= .413107-004
U= .150+001	F(U)= .105917-005	G(X)= .342995-004

.000	.000000	.000000	.775+000	109078-002	125666-001
.250-001	.336041+001	.105735+001	800+000	.828939-003	100448-001
.500-001	.686483+001	.321111+001	[illegible]25+000	6319[illegible]-003	.304248-[illegible]02
.750-001	.722893+001	.455085+001	.850+000	.483291-003	.644984-002
.100+000	.612100+001	.492349+001	.875+000	.370707-003	.518098-002
.125+000	.470783+001	.467910+001	.900+000	.285181-003	416840-002
.150+000	.345121+001	.413799+001	.925+000	.220010-003	.335901-002
.175+000	.246709+001	.350369+001	.950+000	.170201-003	.271099-002
.200+000	.174066+101	.288566+001	.975+000	132022-003	.219134-002
.225+000	.122053+001	.233355+001	.100+001	.102674-003	177396-002
.250+000	.854046+000	.186422+001	.102+001	.800521-004	143821-002
.275+000	.597888+000	.147708+001	.105+001	.625686-004	.116770-002
.300+000	.419435+000	.116392+001	.107+001	.490209-004	.949436-003
.325+000	.295162+000	.913881+000	.110+001	.384964-004	.773054-003
.350+000	.208402+000	.715980+000	.112+001	.303003-004	.630312-003
.375+000	.147886+000	.560267+000	.115+001	.239020-004	.514626-003
.400+000	.105359+000	.438223+000	117+001	.188955-004	4[illegible]8[illegible]55-003
.425+000	.754015-001	.342802+000	.120+001	.149691-004	.344427-003
.450+000	.542088-001	.268300+000	.122+001	.118830-004	.282323-003
.475+000	.391504-001	.210167+000	.125+001	.945195-005	.231711-003
.500+000	.284028-001	.164808+000	.127+001	.753296-005	.190409-003
.525+000	.206973-001	.129402+000	.130+001	.601502-005	.156662-003
.550+000	.151480-001	.101744+000	.132+001	.481189-005	.129051-003
.575+000	.111337-001	.801179-001	.135+001	.385641-005	.106432-003
.600+000	.821729-002	.631876-001	.137+001	.309614-005	.878806-004
.625+000	.608932-002	.499159-001	.140+001	.249006-005	.726456-004
.650+000	.453015-002	.394973-001	.142+001	.200602-005	.601195-004
.675+000	.338309-002	.313059-001	.145+001	.161875-005	.498084-004
.700+000	.253585-002	.248553-001	.147+001	.130837-005	.413107-004
.725+000	.190763-002	.197675-001	.150+001	.105917-005	342995-004
.750+000	.144006-002	.157477-001			



4. DETECCIÓN DE OUTLIERS

Mediante el teorema 2.1 hemos obtenido la función de densidad de el cuadrado de la distancia entre la matriz de suma de cuadrados y suma de productos de una muestra aleatoria de tamaño n extraida de una población $N_2(\mu, \Sigma)$ no singular y la matriz de s. c y s. p. de $n - k$ de estas observaciones.

Debido a que se pueden extraer de $\binom{n}{k}$ formas diferentes estas $n - k$ observaciones de entre las n de la muestra para dada k fijo variando dentro de las condiciones del teorema 2.1. Por otro lado bajo la hipótesis de que los elementos de la muestra de tamaño n pertenecen a la misma población $N_2(\mu, \Sigma)$ se tendrá que $\exists\ q \in \mathbb{R}\ q > 0$ tal que:

$$d^2[S_{(n-k)}^{(i)}, S_{(n)}] < q \qquad \forall i, i = 1, 2, \ldots, \binom{n}{k}$$

y por tanto

$$\max_i d^2[S_{(n-k)}^{(i)}, S_{(n)}] \leqslant q$$

Evidentemente si $\max_i d^2[S_{(n-k)}^{(i)}, S_{(n)}] > q$ esto se interpretaria afirmando que los n elementos que componen la muestra no pertenecen a la misma población. Además los k elementos que se han suprimido de la muestra para obtener $S_{(n-k)}$ serían los elementos que no pertenecen a la población $N_2(\mu, \Sigma)$ ya que al añadir estos k elementos a la muestra de tamaño $n - k$ se produce una dispersión que hace que la distancia entre dichas matrices sobrepase a esa cota. Estos k elementos serán los posibles outliers para esa distribución $N_2(\mu, \Sigma)$.

Al no ser los estadísticos $\{d^2(S_{(n-k)}^{(i)}, S_{(n)});\ i = 1, 2, \ldots, \binom{n}{k}\}$ independientes, vamos a recurrir a hallar una cota superior para la cola de la distribución del estadístico $\max_i d^2(S_{(n-k)}^{(i)}, S_{(n)})$. Es decir acotaremos

$$P\left[\max_i d^2(S_{(n-k)}^{(i)}, S_{(n)}) > q\right]$$

Si definimos C_i como el suceso $\{d^2(S_{(n-k)}^{(i)}, S_{(n)}) > q\}$ para $i = 1, 2, \ldots, \binom{n}{k}$, entonces

$$P\left[\max_i d^2(S_{(n-k)}^{(i)}, S_{(n)}) > q\right] = P\left[\bigcup_{i=1}^{\binom{n}{k}} C_i\right] \leqslant \sum_{i=1}^{\binom{n}{k}} P(C_i) =$$

$$= \binom{n}{k} P(C_i) = \binom{n}{k} P[d^2(S_{(n-k)}^{(i)}, S_{(n)}) > q]$$

sin embargo interesa trabajar con la variable aleatoria X cuya función de densidad viene dada en el teorema 3.2. Por lo que tendremos.

$$P\left[\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q_a\right] \leqslant \binom{n}{k} P\left[d^2(S_{(n-k)}, S_{(n)}) > q_\alpha\right] \leqslant$$

$$\leqslant \binom{n}{k} P[X > q_\alpha] = \alpha$$

En definitiva el proceso a seguir sería el siguiente:

Dado un nivel de significación α se determina el cuantil q_α mediante la función de distribución de la variable aleatoria X.

A continuación calcularemos el valor del estadístico $\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)})$ y la regla de decisión sería la siguiente:

Si

$$\max_i d^2(S_{(n-k)}, S_{(n)}) > q_\alpha$$

los n elementos que componen la muestra no pertenecen a la misma población, siendo considerados como outliers los k elementos que se han suprimido de la muestra para obtener $S_{(n-k)}$.

SUMMARY

We present in this paper a procedure for the detection of outliers for data taken from bivariant normal populational.

The procedure in question appears as a function of the square of the distance between matrices of sums of squares and sums of products of particular data. Such a distance has obtained from the Maas differential metric form.

BIBLIOGRAFÍA

ANDERSON T. W. (1958). An introduction to multivariate statistical analysis. Wiley

FISHER R. A. (1939). The sampling distribution of some statistics obtained from non-linear equations. Annals of Eugenics Vol. 9 pp. 238-249

HSU P. L. (1939) On the distribution of roots of certain determinantal equations. Annals of Eugenics Vol. 9 pp. 250-258.

Kobayashi S. and Nomizu K. (1963). Foundation of differential geomatry. Vol. 1. Wiley Interscience

Krishnaiah P. R. (1978). Some recent developments on real multivariate distribution. In Developments in Statistics, Krishnaiah (ed.). Academic Press.

Maas H. (1955). Die bestimmung der dirichletreihem mit groseencharakteren zu den modulformen n – ten grades. J. Indian Math. Soc. Vol. 19 pp. 1-23.

Muñoz J. (1980). Algunas técnicas sobre detección de outliers. Public. Universidad de Sevilla.

Roy S. N. (1939). *P*-statistics, or some generalizations in analysis of variance appropriate to multivariate problems. Sankhya Vol. 4 pp. 381-396.

Wilks S. S. (1962). Mathematical statistics. Wiley.