Análisis Multivariante, EJ2

Pérez Efremova, Daniel

Enero 2020

Índice

Índice de Códigos

Ejercicio 1

Si la matriz de covarianzas tiene 3 autovalores iguales, entonces, las direcciones de variabilidad (las 3 componentes principales) aglutinan la misma cantidad de información, esto es equivalente a que todas las componentes tienen la misma variabilidad y además son independientes. Con todo esto, debemos plantear un test de esfericidad para comprobar dicha hipótesis.

Usaremos el test de razón de verosimilitudes:

$$\begin{cases} H_0 : V = \sigma^2 \mathbf{I} \\ H_1 : V \neq \sigma^2 \mathbf{I} \end{cases}$$

El estadístico de contraste es:

$$\lambda = np \, log \, \hat{\sigma}^2 - n \, log \, |S| \sim \chi^2_g$$

donde $\hat{\sigma}^2 = \frac{tr\, S}{p}$ (varianza media) es el estimador MV de σ^2 y $g = \frac{(p+2)(p-1)}{2}$ es la diferencia entre las dimensiones de los espacios de ambas hipótesis.

Rechazaremos H_0 si el pvalor del test $P\{\lambda < \chi^2_{(g;1-\alpha)}\}$ es mucho menor que los niveles de significación usuales.

Vemos en el Código 1 los resultados del test.

Código 1: Test de esfericidad

```
S = matrix(c(3,2,2,2,4,3,2,3,5), byrow=TRUE, nrow=3)

p=3
n=70
g = (p+2)*(p-1)/2

sigma_hat = sum(diag(S))/p

lambda = n*p*log(sigma_hat) - n*log(det(S))

pvalue = pchisq(lambda, df=g, lower.tail=FALSE)

cat('estadistico:␣', lambda, '\npvalor:␣', pvalue)

estadistico:  78.00525
pvalor:  2.192672e-15
```

y podemos observar que el valor del estadístico es extremadamente alto como para ser comparado con los percentiles de una χ^2_5, que tiene por esperanza sus grados de libertad (5), mirando el pvalor confirmamos nuestra sospecha de que debemos rechazar la hipótesis nula puesto que es muy cercano a cero.

En conclusión, **deberíamos rechazar que los autovalores de la matriz de covarianzas coinciden**.

Ejercicio 2

apartado a

Para este apartado efectuamos el test de igualdad de medias basado en la razón de verosimilitudes:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$
$$H_1 : \text{no todas las medias son iguales}$$

El estadístico de contraste es $\lambda = n\, log\, \frac{|S|}{|S_w|} \sim \chi^2_g$, donde S es la covarianza sin considerar grupos, $S_w = \frac{1}{n}\sum_{g\in G} S_g$ la covarianza considerando grupos y $g = dim(\Omega_0) - dim(\Omega_1) = p(G-1)$ la diferencia entre las dimensiones de los espacios de ambas hipótesis.

Debemos rechazar H_0 si el pvalor del test $P\{\lambda < \chi^2_{g,1-\alpha}\}$ es suficientemente pequeño. Vemos en el Código 2 el resultado del test.

Código 2: Contraste de igualdad de medias

```
g1 = matrix(c(2,7,5,4,4,5,1,5,4,7,2,3,6,8,2,5,4,3,5,1)
    , byrow=TRUE, nrow=10)
g2 = matrix(c(4,8,5,5,4,8,7,5,4,6,6,8,3,8,6,5,6,7,7,4)
    , byrow=TRUE, nrow=10)

G = rbind(g1,g2)
p=2
n=dim(G)[1]
g = p*(dim(G)[2]-1)

S_w = (cov(g1) + cov(g2))/2
S = cov(G)

lambda = n*log(det(S)/det(S_w))
pvalue = pchisq(lambda, df=g, lower.tail=FALSE)

cat('estadistico:␣',lambda,
'\npvalor:␣', pvalue)

estadistico:  8.637043
pvalor:  0.01331957
```

Y observamos que el pvalor es suficientemente pequeño como para rechazar la igualdad de medias con los niveles de significación usuales $\alpha = 0{,}05$ y $\alpha = 0{,}1$, aun que el valor del estadístico no es excesivamente grande y el pvalor es cercano a 0,01, por lo que no deberíamos rechazar H_0 si exigiéramos gran seguridad en el contraste ($\alpha \leq 0{,}01$).

apartado b

Siguiendo las recomendaciones del texto base, dado que desconocemos las poblaciones, comprobaremos las hipótesis básicas para poder calcular la función discriminante de Fisher: normalidad multivariante en cada población, las poblaciones tienen medias distintas (apartado a) y las poblaciones tienen la misma matriz de covarianzas. Comprobaremos primero si podemos asumir la misma matriz de covarianzas para los grupos con el test de la M de Box.

```
library(covTestR)
homogeneityCovariances(list(g1,g2))

Boxes' M Homogeneity of Covariance Matrices Test

data:
Chi-Squared = 4.7989, df = 55, p-value = 1
alternative hypothesis: true difference in covariance
   matrices is not equal to 0
```

El test arroja un pvalor alto por lo que no debemos rechazar que las poblaciones tienen la misma matriz de covarianzas.

Por otro lado la normalidad multivariante en cada uno de los grupos la comprobaremos con el test de Mardia.

```
library(MVN)

normality_test1 = mvn(g1,
cov = TRUE, mvnTest='mardia')
normality_test2 = mvn(g2,
cov = TRUE, mvnTest='mardia')
cat('test de Mardia P1')
cat('\n\n')
print(normality_test1.multivariateNormality)
cat('\n\n')

cat('test de Mardia P2')
cat('\n\n')
print(normality_test2.multivariateNormality)
cat('\n\n')
```

```
test de Mardia P1

Test           Statistic            p value Result
1 Mardia Skewness   3.56517825480128 0.468037296747371
       YES
2 Mardia Kurtosis -0.689148747967349  0.49072966570324
       YES
3              MVN                <NA>               <NA>
       YES


test de Mardia P2

Test          Statistic            p value Result
1 Mardia Skewness 0.838465309927981 0.933218128142171
      YES
2 Mardia Kurtosis  -1.2659175760541 0.205542593128733
      YES
3              MVN               <NA>               <NA>
      YES
```

Observamos pvalores altos en cuanto a kurtosis y simetría, por lo que no deberíamos rechazar la normalidad multivariante. Aun que el test tiene poca potencia debido al tamaño reducido de las muestras, asumiremos normalidad para completar el ejercicio.

En las condiciones comprobadas anteriormente, la función discriminante de Fisher para el grupo g es:

$$L_g(x) = w_g' \bar{x}_g - 2w_g' x \tag{1}$$

donde $w = \hat{S}^{-1}\bar{x}_g$. Vemos en el Código 3 los cálculos para obtener la función discriminante:

Código 3: Cálculo de la función discriminante lineal de Fisher

```
library(matlib)
mean1 = colMeans(g1)
mean2 = colMeans(g2)

w1 = inv(S)%*%mean1
w2 = inv(S)%*%mean2

cat('w1␣:', w1)
cat('\nw2␣:', w2)

w1 =  1.159322, 1.175424
```

```
w2 =  1.724982, 1.565365
```

Con estos resultados las funciones discriminantes se deducen de (1) y para un punto (x_1, x_2) son:

$$L_1(x) = 9{,}7 - 2{,}31x_1 - 2{,}35x_2$$

$$L_2(x) = 18{,}9 - 3{,}44x_1 - 3,13x_2$$

apartado d

Para clasificar la observación $(2, 1)$ en una de las dos poblaciones usamos la regla $\text{mín}_{g \in G} L_g(x)$ entre las funciones calculadas en el apartado anterior. Vemos en el código 4 el cálculo de ambas.

Código 4: Clasificación de la observación (2,1)

```
l1 = function(x){
        9.7 -2.31*x[1] - 2.35*x[2]
}

l2 = function(x){
        18.7 -3.44*x[1] - 3.13*x[2]
}

cat('L1:', l1(c(1,2)), '\nL2:', l2(c(1,2)))

L1: 2.69
L2: 9
```

Observamos que el valor mínimo se encuentra en el grupo 1, por lo que **la nueva observación debe ser considerada del grupo 1**.

apartado d

Como se aconseja en el texto base, cuando desconozcamos las poblaciones debemos construir la probabilidad de error a través de la tabla de contingencia (*confusion matrix*) entre los grupos a los que pertenecen originalmente las observaciones y los grupos asignados por la regla de decisión.

Así, la probabilidad de error vendrá dada por

$$p_{error} = \frac{\text{total mal clasificados}}{\text{total bien clasificados}}$$

Vemos en el código 4 la tabla de contingencia.

Código 5: Cálculo de las probabilidades de error. Poblaciones desconocidas.

```
#definimos un dataframe para almacenar las
    observaciones y el grupo
grupo = append(rep(1,10), rep(2,10))  # columna de
    grupos
G = cbind(G, grupo) # nueva matriz de datos
G = data.frame(G)

clasificador = function(x){

        grupos = c('1', '2')
        valores_dicrim = c(l1(x), l2(x))
        return (grupos[which.min(valores_dicrim)])
}

pred = apply(G[, c(1,2)], 1, clasificador)
G.pred = pred  # columna de predicciones

library(caret)
confusionMatrix(factor(G.grupo), factor(G.pred))

Confusion Matrix and Statistics

             Reference
Prediction   1  2
         1   8  2
         2   0 10

Accuracy : 0.9
95% CI : (0.683, 0.9877)
```

Y de la matriz de confusión deducimos que la probabilidad de error aproximada es:

$$p_{error} = \frac{2}{18} = \frac{1}{9} \approx 0{,}11$$

por otro lado, podemos ver que la precisión es aproximadamente de 0,9 y se encuentra en el intervalo de confianza $I = (0{,}68, 0{,}98)$ al 95 %.