Zadatak 51. Test jednakosti kovarijanci

Podaci: ES, Table 14.10. str. 551.
(a) Opišite test omjera vjerodostojnosti za hipotezu o jednakosti kovarijacijskih matrica (TMS 8.7, str. 121. - 124.) normalnih uzoraka.
(b) Ispitajte normalnost podataka.
(c) Sprovedite test iz (a) za podatke iz zadatka, tj. usporedite kovarijance grupa 1 i 2.

Rješenje

a) Test omjera vjerodostojnosti

Neka je zadano a međusobno nezavisnih uzoraka $x_{i1}, \dots, x_{in_i}$, svaki s distribucijom $N_p(\mu_i, \Sigma_i)$, gdje je $\Sigma_i > 0$, $i = 1, \dots, a$. Označimo s $n = \sum_{i=1}^{a} n_i$ ukupan broj opažanja. Želimo testirati:

$$H_0 : \Sigma_1 = \cdots = \Sigma_a$$
$$H_1 : ne - H_0$$

Funkcija vjerodostojnosti je uobičajena (bez konstanti koje će se pokratiti):

$$L(\Sigma_1, \dots, \Sigma_a, \mu_1, \dots, \mu_a) \propto \prod_{i=1}^{a} |\Sigma_i|^{-\frac{n_i}{2}} \operatorname{etr}\{-\frac{1}{2}[V_i + n_i(\bar{x}_i - \mu_i)(\bar{x}_i - \mu_i)']\Sigma_i{}^{-1}\},$$

uz oznake:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij},$$
$$V_i = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)', i = 1, \dots, a$$

te uz pokratu $\operatorname{etr}(A) = \exp(\operatorname{tr}(A))$, za neku matricu A. Znamo da su MLE procjenitelji za μ_i i Σ_i jednaki $\hat{\mu}_i = \bar{x}_i$, tj. $\hat{\Sigma}_i = \frac{1}{n_i}V_i$. Ako je ispunjena H_0, imamo $\Sigma_1 = \cdots = \Sigma_a = \Sigma$, za neki Σ. Iz toga slijedi da je $\hat{\hat{\mu}}_i = \bar{x}_i$ i $\hat{\hat{\Sigma}}_i = \frac{1}{n}V$, gdje je

$V = \sum_{i=1}^{a} V_i$. Dobivamo da je testna statistika za test omjera vjerodostojnosti:

$$\begin{aligned}\Lambda &= \frac{L(\frac{1}{n}V, \dots, \frac{1}{n}V, \bar{x}_1, \dots, \bar{x}_a)}{L(\frac{1}{n_1}V_1, \dots, \frac{1}{n_a}V_a, \bar{x}_1, \dots, \bar{x}_a)} \\ &= \frac{\prod_{i=1}^{a} \left|\frac{1}{n}V\right|^{-n_i/2} \exp(-\frac{1}{2}np))}{\prod_{i=1}^{a} \left|\frac{1}{n_i}V_i\right|^{-n_i/2} \exp(-\frac{1}{2}n_i p))} \\ &= \frac{\prod_{i=1}^{a} |V_i|^{n_i/2}\, n^{pn/2}}{|V|^{n/2} \prod_{i=1}^{a} n_i{}^{pn_i/2}}\end{aligned}$$

Direktna posljedica toga (i činjenice da znamo kakvog je oblika kritično područje za test omjera vjerodostojnosti) je sljedeća propozicija:

Propozicija 1 *Test omjera vjerodostojnosti za testiranje* $H_0 : \Sigma_1 = \dots = \Sigma_a$ *odbacuje nultu hipotezu u korist alternativne za male vrijednosti*

$$\Lambda = \frac{\prod_{i=1}^{a} |V_i|^{n_i/2}\, n^{pn/2}}{|V| \prod_{i=1}^{a} n_i{}^{pn_i/2}}.$$

Primijetimo da općenito ne znamo distribuciju naše testne statistike, međutim možemo ju aproksimirati. Ako testiramo:

$$\begin{aligned}H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_0^{\complement}\end{aligned}$$

te s Λ označimo testnu statistiku za test omjera vjerodostojnosti (tj. $\Lambda(x) = \frac{sup\{L(\theta|x):\theta\in\Theta_0)\}}{sup\{L(\theta|x):\theta\in\Theta)\}}$), tada $-2log(\Lambda)$ ima asimptotsku χ^2 distribuciju s d stupnjeva slobode, gdje je d razlika dimenzija Θ i Θ_0. Također i iz toga vidimo da za male vrijednosti Λ odbacujemo nultu hipotezu jer je u tom slučaju $-2log(\Lambda)$ jako veliko, a samim time p-vrijednost mala.

Promatrajmo sad transformirane podatke $x_{ij} \mapsto Ax_{ij} + b_i$, gdje je $A \in G_p$, gdje je G_p grupa simetričnih matrica reda p, a $b_i \in \mathbb{R}^p, i = 1, \dots, a$. Znamo da ovakva tranformacija čuva normalnost te da $\bar{x}_i \mapsto A\bar{x}_i + b_i$, a $V_i \mapsto AV_iA'$, odnosno $V \mapsto AVA'$, što nam sugerira $\Sigma_i \mapsto A\Sigma_iA'$. Dakle, ovakvom transformacijom se čuva nulta hipoteza, tj. iznosi $H_0 : A\Sigma_1A' = \dots = A\Sigma_aA'$. Sada znamo da testna statistika za tranformirane podatke iznosi:

$$\Lambda = \frac{\prod_{i=1}^{a} |AV_iA'|^{n_i/2}\, n^{pn/2}}{|AVA'| \prod_{i=1}^{a} n_i{}^{pn_i/2}} = \frac{\prod_{i=1}^{a} |V_i|^{n_i/2}\, n^{pn/2}}{|V| \prod_{i=1}^{a} n_i{}^{pn_i/2}},$$

Što je jednako testnoj statistici za netransformirane podatke. Kažemo da je test omjera vjerodostojnosti invarijantan obzirom na ovakve transformacije. Općenito, kažemo da je testna funkcija $f(\bar{x}_1, \ldots, \bar{x}_a, V_1, \ldots, V_a)$ *invarijantna* ako

$$\begin{aligned}&f(y_1, \ldots, y_a, W_1, \ldots, W_a) = f(Ay_1 + b_1, \ldots, Ay_a + b_a, AW_1A', \ldots, AW_aA'),\\ &\forall(A, b_1, \ldots, b_a) \in G_p \times (\mathbb{R}^p)^a, \forall(y_1, \ldots, y_a, W_1, \ldots, W_a) \in (\mathbb{R}^p)^a \times (\mathcal{P}_p)^a.\end{aligned}$$

Invarijantne testne funkcije imaju dva zanimljiva svojstva.

Prvo, ako matricu A definiramo kao $A = \Sigma^{-1/2}$, gdje je uz H_0 $\Sigma_1 = \cdots = \Sigma_a = \Sigma$ te stavimo $B_i = -A\bar{x}_i$, dobivamo da je $AV_iA' \sim W_p(n_i - 1)$, a to ne uključuje nikakve parametre, odnosno:

$$\begin{aligned}&f(\bar{x}_1, \ldots, \bar{x}_a, V_1, \ldots, V_a)\\ &= f(A\bar{x}_1 + b_1, \ldots, A\bar{x}_a + b_a, AV_1A', \ldots, AV_aA')\\ &= f(0, \ldots, 0, AV_1A', \ldots, AV_aA').\end{aligned}$$

Primijetimo da trebamo samo uzimati u obzir testove tog oblika, s obzirom da je $(\bar{x}_1, \ldots, \bar{x}_a, V_1, \ldots, V_a)$ dovoljno za $(\mu_1, \ldots, \mu_a, \Sigma_1, \ldots, \Sigma_a)$.

Drugo zanimljivo svojstvo dobivamo u slučaju kad nam je $a = 2$, dijagonalizirajući matricu na sljedeći način: $V_1^{-1/2}V_2V_1^{-1/2} = HDH'$, gdje je $H \in O_p$, a $D = \operatorname{diag}(l_1, \ldots, l_p)$ dijagonalna matrica koja sadrži svojstvene vrijednosti od $V_1^{-1}V_2$. Uzevši sada $A = H'V_1^{-1/2}$, $b_i = -Ax_i$, dobivamo da je za svaki invarijantni test

$$\begin{aligned}f(\bar{x}_1, \bar{x}_2, V_1, V_2) &= f(0, 0, AV_1A', AV_2A')\\ &= f(0, 0, I, D)\end{aligned}$$

funkcija samo svojstvenih vrijednosti $l_1, \ldots, l_p$, odnosno svaki invarijantni test ovisi o podacima samo preko svojstvenih vrijednosti. Slično, možemo dijagonalizirati $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} = GD_\lambda G'$, gdje je $G \in G_p$, a D_λ sadrži svojstvene vrijednosti $\lambda_1, \ldots, \lambda_p$ matrice $\Sigma_1^{-1}\Sigma_2$. Uzevši sada $A = G'\Sigma_1^{-1/2}$, dobivamo $AV_1A' \sim W_p(n_1 - 1)$ i $AV_2A' \sim W_p(n_i - 1)$. To je sadržaj sljedeće propozicije:

Propozicija 2 *Obzirom na grupu transformacija $G_p \times (\mathbb{R}^p)^2$, bilo koji invarijantni test za testiranje $H_0 : \Sigma_1 = \Sigma_2$ ovisi o $(\bar{x}_1, \bar{x}_2, V_1, V_2)$ samo preko svojstvenih vrijednosti $l_1, \ldots, l_p$ matrice $V_1^{-1}V_2$. Funkcija jakosti testa bilo kojeg invarijantnog testa ovisi o $(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$ samo preko svojstvenih vrijednosti $\lambda_1, \ldots, \lambda_p$ od $\Sigma_1^{-1}\Sigma_2$.*

b) Normalnost podataka

Dani su podaci o dvije vrste klokana:

1. M. giganteus (označena s **M**)
2. M.f. melanops (označena sa **Z**)

Stupci danih podataka predstavljaju redom: okonosnu duljinu, nosnu duljinu, nosnu širinu, zigomatičnu širinu, širinu grebenu, širinu mandibule, uzlaznu visinu ramusu. Želimo testirati normalnost podataka pomoću Lillieforsove inačice Kolmogorov - Smirnovljevog testa, gdje nam je nulta hipoteza da su podaci normalno distribuirani. Test provodimo u R-u. Zadani su podaci:

$$\mathbf{M} = \begin{pmatrix} 1312 & 1445 & 609 & 241 & 782 & 153 & 179 & 591 \\ 1439 & 1503 & 629 & 222 & 824 & 141 & 181 & 643 \\ 1378 & 1464 & 620 & 233 & 778 & 144 & 169 & 610 \\ 1315 & 1367 & 564 & 207 & 801 & 116 & 189 & 594 \\ 1413 & 1500 & 645 & 247 & 823 & 120 & 197 & 654 \\ 1090 & 1195 & 493 & 189 & 673 & 188 & 138 & 476 \\ 1294 & 1421 & 606 & 226 & 780 & 149 & 168 & 578 \\ 1377 & 1504 & 660 & 240 & 812 & 128 & 175 & 628 \\ 1296 & 1439 & 630 & 215 & 759 & 151 & 159 & 578 \\ 1470 & 1563 & 672 & 231 & 856 & 103 & 196 & 683 \\ 1612 & 1699 & 778 & 263 & 921 & 86 & 232 & 772 \\ 1388 & 1500 & 616 & 220 & 805 & 107 & 180 & 652 \\ 1575 & 1655 & 727 & 271 & 905 & 82 & 210 & 712 \\ 1717 & 1821 & 810 & 284 & 960 & 104 & 222 & 731 \\ 1587 & 1711 & 778 & 279 & 910 & 81 & 207 & 692 \\ 1604 & 1770 & 823 & 272 & 880 & 57 & 208 & 713 \\ 1603 & 1703 & 755 & 268 & 902 & 81 & 206 & 754 \\ 1490 & 1599 & 710 & 278 & 897 & 115 & 194 & 688 \\ 1552 & 1540 & 701 & 238 & 852 & 82 & 213 & 722 \\ 1595 & 1709 & 803 & 255 & 904 & 83 & 183 & 701 \\ 1840 & 1907 & 855 & 308 & 984 & 84 & 238 & 795 \\ 1740 & 1817 & 838 & 281 & 977 & 121 & 227 & 770 \\ 1846 & 1893 & 830 & 288 & 1013 & 21 & 232 & 829 \\ 1702 & 1860 & 864 & 306 & 947 & 39 & 218 & 776 \\ 1768 & 1890 & 837 & 285 & 968 & 41 & 243 & 842 \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix}
1299 & 1345 & 565 & 204 & 764 & 153 & 156 & 556 \\
1337 & 1395 & 562 & 216 & 794 & 154 & 158 & 625 \\
1372 & 1456 & 580 & 225 & 814 & 124 & 179 & 636 \\
1336 & 1441 & 596 & 220 & 788 & 156 & 178 & 623 \\
1301 & 1387 & 579 & 219 & 787 & 113 & 164 & 616 \\
1360 & 1467 & 636 & 201 & 813 & 138 & 171 & 603 \\
1276 & 1351 & 559 & 213 & 766 & 129 & 159 & 608 \\
1613 & 1726 & 740 & 234 & 883 & 75 & 184 & 745 \\
1542 & 1628 & 677 & 237 & 885 & 94 & 190 & 709 \\
1440 & 1580 & 675 & 217 & 815 & 129 & 186 & 634 \\
1474 & 1555 & 629 & 211 & 888 & 134 & 205 & 716 \\
1503 & 1603 & 692 & 238 & 825 & 83 & 203 & 712 \\
1597 & 1653 & 710 & 221 & 908 & 104 & 194 & 761 \\
1671 & 1689 & 730 & 281 & 892 & 62 & 208 & 770 \\
1673 & 1720 & 763 & 292 & 946 & 107 & 196 & 755 \\
1458 & 1588 & 686 & 251 & 836 & 115 & 192 & 676 \\
1568 & 1689 & 717 & 231 & 900 & 18 & 194 & 759 \\
1650 & 1707 & 737 & 275 & 943 & 72 & 184 & 768 \\
1774 & 1838 & 816 & 275 & 994 & 56 & 227 & 794 \\
1893 & 1945 & 893 & 260 & 994 & 13 & 216 & 824 \\
1765 & 1781 & 766 & 261 & 978 & 38 & 211 & 775
\end{pmatrix}$$

Provođenje Lillieforsove inačice Kolmogorov - Smirnovljevog testa:

```
> M <- read.csv("muski.txt",header=FALSE, sep = ',')
> x <- c(1,2,3,4,5,6,7,8)
> for (val in x) {print(lillie.test(M[,val]))}

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.092918, p-value = 0.8355

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.12852, p-value = 0.3558
```

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.12962, p-value = 0.3428

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.11233, p-value = 0.5733

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.11264, p-value = 0.5689

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.13076, p-value = 0.3295

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.089923, p-value = 0.8676

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  M[, val]
D = 0.080297, p-value = 0.9463
```

```
> Z <- read.csv("zenski.txt",header=FALSE, sep = ',')
> for (val in x) {print(lillie.test(Z[,val]))}

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.13018, p-value = 0.465

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.11735, p-value = 0.6327

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.11305, p-value = 0.6893

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.15545, p-value = 0.2049

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.14347, p-value = 0.3122

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.1227, p-value = 0.5617

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Z[, val]
D = 0.082292, p-value = 0.9695

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  Z[, val]
D = 0.17214, p-value = 0.107
```

Pošto su sve p-vrijednosti veće od 0.05, ne odbacujemo nultu hipotezu, tj. svi su uzorci normalno distribuirani.

c) Test omjera vjerodostojnosti

Izračunajmo najprije kovarijacijske matrice:

```
> KovM = cov (M)
> KovM
  V1        V2        V3        V4        V5         V6        V7
36582.193 35425.917 19082.443 5484.8900 15766.810 -6403.3017 4735.0133
35425.917 35734.667 19394.083 5607.5417 15335.875 -6317.3333 4483.6250
19082.443 19394.083 10907.443 3076.8483  8253.060 -3429.7183 2386.6800
5484.890  5607.542  3076.848 1000.6933  2433.940  -948.6567  708.4450
15766.810 15335.875  8253.060 2433.9400  7064.177 -2751.6683 2075.3633
-6403.302 -6317.333 -3429.718 -948.6567 -2751.668  1604.9100 -863.2550
4735.013  4483.625  2386.680  708.4450  2075.363  -863.2550  710.5067
16330.455 15746.792  8446.288 2410.6283  7081.472 -3100.5300 2231.4150

 V8
16330.455
15746.792
8446.288
2410.628
7081.472
-3100.530
2231.415
7786.740
```

```
> KovZ = cov (Z)
> KovZ
 V1        V2        V3        V4        V5         V6         V7
31398.329 28865.700 15523.200 3791.2143 12716.607 -6460.8143 2989.9000
28865.700 27507.233 14774.283 3365.6167 11637.500 -6103.4500 2845.0833
15523.200 14774.283  8176.433 1845.1167  6146.400 -3258.6000 1490.8333
3791.214  3365.617  1845.117  734.2905  1528.879  -734.6571  354.8667
12716.607 11637.500  6146.400 1528.8786  5547.214 -2537.3786 1226.5500
-6460.814 -6103.450 -3258.600 -734.6571 -2537.379  1868.8571 -605.2500
2989.900  2845.083  1490.833  354.8667  1226.550  -605.2500  387.4333
12979.550 12066.233  6299.783 1582.6167  5382.700 -2824.5500 1301.2833

V8
12979.550
12066.233
6299.783
1582.617
5382.700
-2824.550
1301.283
5914.133

> lM = length(M[,1])
> lM
[1] 25
> lZ = length(Z[,1])
> lZ
[1] 21
```

Izračunajmo V-ove:

```
> VM = (lM - 1)*KovM

> VZ = (lZ - 1)*KovZ

> V = VM + VZ
> V
 V1        V2         V3        V4         V5         V6         V7
1505939.2 1427536.0  768442.64 207461.65  632735.58 -282895.53 173438.32
1427536.0 1407776.7  760943.67 201893.33  600811.00 -273685.00 164508.67
768442.6  760943.7  425307.31 110746.69  321001.44 -147485.24  87096.99
207461.6  201893.3  110746.69  38702.45   88992.13  -37460.90  24100.01
632735.6  600811.0  321001.44  88992.13  280484.53 -116787.61  74339.72
-282895.5 -273685.0 -147485.24 -37460.90 -116787.61   75894.98 -32823.12
173438.3  164508.7   87096.99  24100.01   74339.72  -32823.12  24800.83
651521.9  619247.7  328706.59  89507.41  277609.32 -130903.72  79579.63

 V8
651521.92
619247.67
328706.59
89507.41
277609.32
-130903.72
79579.63
305164.43
```

Izračunajmo vrijednost testne statistike Λ:

```
det(VM)^(25/2)*det(VZ)^(21/2)/det(V)^(46/2)*46^(8*46/2)/(25^(8*25/2)*21^(8*21/2))

  2.50922e-009
```

Dakle, Λ je jako mali, pa prema **Propoziciji 1** odbacujemo nultu hipotezu, tj. zaključujemo da kovarijacijske matrice nisu jednake.

Pretpostavimo sada da $-2log(\Lambda)$ ima aproksimativno normalnu razdiobu:

```
t=-2*log(L)
```

```
t =
  17.20092

d=36



1-chi2cdf(t,d)

ans =
  0.003383737
```

Kako nam je p-vrijednost mala, odbacujemo hipotezu o jednakosti kovarijacijskih matrica.