MANOVA

Tulajdonságok:

- Hasonló az ANOVÁ-hoz
- Több függő változó
- A függő változók korreláltak és a lineáris kombinációnak értelme van.
- Azt teszteli, hogy *k* populációban a függő változók egy lineáris kombinációjának átlagai különböznek-e.

Alapötlet: találjunk egy olyan lineáris kombinációt, amely optimálisan szeparálja a csoportokat, azaz olyat amely maximalizálja a hatás (between group) variancia/kovariancia mátrix és a hiba (within group) variancia/kovariancia mátrix hányadosát. (Ez ugyanaz, mint amit a diszkriminancia elemzésnél használunk.)

Ennek a kombinációnak a standardizált együtthatói megmondják, hogy melyik változó milyen súllyal szerepel a szeparálásban.

Előnyök:

- Annak az esélye, hogy különbségeket találunk a csoportok között nagyobb ahhoz képest, mintha minden változóra egyenként ANOVÁ-t csinálnánk.
- Nem inflálódik az elsőfajú hiba.
- Több ANOVA elvégzése nem veszi figyelembe azt, hogy a függő változók korreláltak.

Hátrányok:

- Bonyolultabb,
- Az ANOVA gyakran nagyobb hatóerejű.
- Sokkal komplikáltabb kísérleti elrendezést igényel.
- Kétségek merülhetnek fel, hogy valójában mely független változók mely függő változók értékét befolyásolják.
- Minden plusz függő változó 1 szabadsági fokkal kevesebbet jelent.

Feltételek:

Független minták,

Többváltozós normális eloszlású hiba.

A kovariancia mátrix homogenitása.

Lineáris kapcsolat a független változók és a függő változók között.

A MANOVA elvégzésének lépései:

Ha a MANOVA nem szignifikáns, stop

Ha a MANOVA szignifikáns, egyváltozós ANOVÁk

Ha az egyváltozós ANOVA szignifikáns, Post Hoc tesztek.

Ha igaz a homoscedasticity, Wilks Lambda, ha nem Pillai's Trace. Általában mind a 4 statisztikának hasonlónak kell lennie.

A MANOVA algoritmusa:

1. Az ANOVA négyzetösszegei helyett sums-of-squares-and-cross-products (SSCP) mátrixok. Egy a hatásnak (between

2. Kiszámítjuk a $\mathbf{HE}^{-1}$ szorzatot (egyváltozós esetben ez az F érték).
3. Kiszámítjuk a $\mathbf{HE}^{-1}$ spektrál felbontását: sajátértékek, sajátvektorok. A s.é.-kek azt mutatják meg, hogy between-group varianciából a sajátvektorok vagy lineáris kombinációk mennyit magyaráznak. A s.v.-ok tartalmazzák a lineáris kombinációk együtthatóit.
4. Az a lineáris kombináció, amelyikhez a legnagyobb s.é. tartozik maximalizálja a between-group/within-group variancia hányadost.

H_0: a csoport centroidok megegyeznek.

Ez tesztelhető valamelyik variancia mérték segítségével (nyom, determináns: általánosított variancia).

- Wilk's lambda: $|\mathbf{E}|/|\mathbf{T}|$. A teljes variancia hányad része a reziduális. Minél kisebb, annál nagyobb a csoportok köztötti különbségek.
- Hotelling-Lawley trace: $|\mathbf{H}|/|\mathbf{E}|$. Ez ugyanaz, mint a $\mathbf{HE}^{-1}$ mátrix nyoma (sajátértékek összege). Nagyobb értékek nagyobb különbségeket indikálnak a csoport centroidok között.
- Pillai trace: A $\mathbf{HT}^{-1}$ nyoma, vagyis a between groups variancia.
- Roy's largest root: a $\mathbf{HE}^{-1}$ legnagyobb s.é.-e, vagyis ahhoz a lineáris kombinációhoz tartozó s.é. amely a between groups variancia-kovarianca legnagyobb részét magyarázza.

Ezeknek a statisztikáknak az eloszlása nem teljesen ismert, közelítő F értékekké konvertálják ezeket. Két csoport esetén a Wilk's lambda, a Hotteling és Pillai féle érték megegyezik és megegyezik a Hotteling féle T^2 statisztikával, ami a *t*-próba többváltozós kiterjesztése. Általában hasonló eredményeket produkálnak több csoport esetén is. A Pillai trace a legrobosztusabb teszt.

```
> attach(skulls)
> skulls.manova<-manova(cbind(MB,BH,BL,NH)~EPOCH)
> summary(skulls.manova,test="Pillai")
           Df Pillai approx F num Df den Df    Pr(>F)
EPOCH       4 0.3533   3.5120     16    580 4.675e-06 ***
Residuals 145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(skulls.manova,test="Wilks")
              Df  Wilks approx F num Df den Df   Pr(>F)
EPOCH       4.00 0.6636   3.9009  16.00 434.45 7.01e-07 ***
Residuals 145.00
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(skulls.manova,test="Hotelling")
           Df Hotelling-Lawley approx F num Df den Df
EPOCH       4           0.4818   4.2310     16    562
Residuals 145
             Pr(>F)
EPOCH     8.278e-08 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(skulls.manova,test="Roy")
           Df     Roy approx F num Df den Df    Pr(>F)
EPOCH       4  0.4251  15.4097      4    145 1.588e-10 ***
```

Diszkriminancia analízis

Cél: egy olyan függvény létrehozása, amely alapján az egyedek két vagy több csoportba sorolhatók (a függvény értéke lényegesen változik csoportról csoportra). Később a függvényt új egyedek besorolására lehessen használni.
pl. verebek. A testméretek alapján besorolhatók-e a verebek a túlélők ill. nem túlélők közé (Mire emlékeztet ez a kérdés?!!):

Lineáris diszkriminancia függvény:

$$Z = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p$$

Ha Z értéke jelentősen változik csoportról csoportra, akkor a csoportok jól szeparálhatók. Több függvény is konstruálható.
A függvény úgy vetíti le a csoportokat egy alacsonyabb dimenziós térbe, hogy azok eloszlásai a legkisebb mértékben fedjék át egymást.

A MANOVA inverze. A MANOVA ugyanezt a függvényt használja.

Kétféle cél:

1. Prediktív diszkriminancia analízis (generáljunk egy szabályt, amely alapján csoportokba sorolhatunk).

2. Leíró analízis: a függő változó és a független változók kapcsolatát vizsgáljuk.

Hogyan működik?

1. Feltételezzük, hogy a célpopuláció egymást kizáró rész populációkból áll.

2. Feltételezzük, hogy a független változóink többváltozós normális elsozlást követnek
3. Megkeressük azt a lineáris kombinációt, amely a legjobban szeparálja a csoportokat.
4. Ha *k* csoportunk van, akkor *k*-1 diszkriminancia függvényt készítünk.
5. Minden függvényre kiszámítjuk a diszkriminancia szkórokat.
6. Ezeket a szkórokat használjuk a klasszifikáláshoz.

```
> skulls.lda<-lda(EPOCH~.,skulls)
> skulls.lda
Call:
lda(EPOCH ~ ., data = skulls)

Prior probabilities of groups:
c1850BC  c200BC c3300BC c4000BC  cAD150
    0.2     0.2     0.2     0.2     0.2

Group means:
               MB       BH       BL       NH
c1850BC 134.4667 133.8000 96.03333 50.56667
c200BC  135.5000 132.3000 94.53333 51.96667
c3300BC 132.3667 132.7000 99.06667 50.23333
c4000BC 131.3667 133.6000 99.16667 50.53333
cAD150  136.1667 130.3333 93.50000 51.36667

Coefficients of linear discriminants:
           LD1         LD2         LD3           LD4
MB  0.12667629 -0.03873784 -0.09276835 -0.1488398644
BH -0.03703209 -0.21009773  0.02456846  0.0004200843
BL -0.14512512  0.06811443 -0.01474860 -0.1325007670
NH  0.08285128  0.07729281  0.29458931 -0.0668588797
```

```
Proportion of trace:
   LD1    LD2    LD3    LD4
0.8823 0.0809 0.0326 0.0042
```

Logisztikus ill. multinomiális regresszió vagy diszkriminancia analízis?

Ha a magyarázó változók normális eloszlásúak, akkor a DA jobb.

Ha kategóriás változóink is vannak, akkor a DA akkor rosszabb, ha a kategóriák száma nagyon kicsi (2, 3). Ezekben az esetekben a LR eredménye hasonló a DA-éhoz, legfeljebb egy kicsit rosszabb (ha a mintaelemszám aránylag kicsi).

Ha a DA feltételei nem teljesülnek, mindenképpen a LR-t kell használni. Az LR nem eloszlás függő.

Kanonikus korreláció elemzés

Többszörös regresszió elemzés általánosítása.

Sokszor két természetes csoportot alkotnak a változók és a két csoport közötti kapcsolatot szereténk vizsgálni.

Példa: 16 Euphydryas editha lepke kolónia Kaliforniából és Oregonból. Minden kolónia esetén ismert 4 környezeti változó és 6 génfrekvencia érték. Kérdés: milyen kapcsolatban vannak egymással a környezeti és genetikus tényezők?

Változók: *Alt – Tengerszint feletti magasság (láb)*

prec- éves csapadék mennyiség

max – Éves max. hőmérséklet (°F)

min – Éves min. hőmérséklet (°F)

F0.40-F1.30 Pgi mobility gene frequencies (%)

Colony	Alt	prec	max	min	F0.40	F0.60	F0.80	F1.00	F1.16	F1.30
SS	500	43	98	17	0	3	22	57	17	1
SB	800	20	92	32	0	16	20	38	13	13
WSB	570	28	98	26	0	6	28	46	17	3
JRC	550	28	98	26	0	4	19	47	27	3
JRH	550	28	98	26	0	1	8	50	35	6
SJ	380	15	99	28	0	2	19	44	32	3
CR	930	21	99	28	0	0	15	50	27	8
UO	650	10	101	27	10	21	40	25	4	0
LO	600	10	101	27	14	26	32	28	0	0
DP	1500	19	99	23	0	1	6	80	12	1
PZ	1750	22	101	27	1	4	34	33	22	6
MC	2000	58	100	18	0	7	14	66	13	0
IF	2500	34	102	16	0	9	15	47	21	8
AF	2000	21	105	20	3	7	17	32	27	14
GH	7850	42	84	5	0	5	7	84	4	0
GL	10500	50	81	-12	0	3	1	92	4	0

Ötlet: Készítsünk olyan lineáris kombinációkat a két csoportban lévő változókból, hogy azok maximálisan korreláltak legyenek.

A gyakorlatban több változó készíthető. Ha van p $(X_1, X_2, ..., X_p)$ és q $(Y_1, Y_2, ..., Y_q)$ **standardizált** változónk a két csoportban, akkor min(p,q) ilyen lineáris kombináció készíthető. Azaz

$$U_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \qquad i = 1, 2, ..., r$$
$$V_i = a_{i1}Y_1 + a_{i2}Y_2 + ... + a_{ip}Y_p$$

ahol $r = \min(p,q)$

Úgy választjuk meg az együtthatókat, hogy az U_1 és V_1 korrelációja maximális legyen, U_2 és V_2 korrelációja maximális legyen olyan módon, hogy nem korreláltak U_1 –gyel és V_1-gyel, stb...
Ilyen módon minden (U_i és V_i) a kapcsolat különböző „dimenzióit” méri. Az első pár korrelációja a legnagyobb.

A számítás menete

Elkészítünk egy (p+q)×(p+q) dimenziós korrelációs mátrixot a változóinkból:

$$\begin{array}{c} \\ X_1 \\ X_p \\ Y_1 \\ Y_q \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & & X_p & Y_1 & Y_2 & & Y_q \end{array} \\ \begin{bmatrix} A & & \vdots & & C \\ \cdots & & & & \\ & & \cdots & & \cdots \\ C^T & & & & \\ & & \vdots & & B \end{bmatrix} \end{array}$$

Kiszámítható egy $\mathbf{B^{-1}C^{T}A^{-1}C}$ mátrix kiszámíthatóak ennek a sajátértékei. Bebizonyítható, hogy a $\lambda_1 > \lambda_2 > ... > \lambda_r$, a kanonikus változók korrelációinak (kanonikus korrelációk) négyzetei. A hozzájuk tartozó $\mathbf{b_1, b_2, ..., b_r}$ sajátvektorok pedig az Y_i-k együtthatói. Az X_i-k együtthatóit az

$$\mathbf{a}_i = \mathbf{A^{-1}Cb_i}$$

vektor komponensei adják.
A sajátértékek azt mutatják meg, hogy a független változók mennyit magyaráznak a függőkből az adott dimenzióban.

Szignifikancia tesztek

Ha *r* sajátértékünk van, akkor *r* kanonikus változó párunk van. Ezek közül sok olyan kicsi, hogy már nem szignifikáns. A Wilk's féle tesztet használjuk annak eldöntésére, hogy hány szignifikáns változó párunk van. A szabadsági foka $p*q$.

Feltételek

Általában ugyanazok mint a MANOVA esetén:

- linearitás
- intervallum vagy legalábbis közel intervallum skálán mért változók
- többváltozós normalitás

Példa

Az utolsó frekvencia változó nem kell, mert a 6 együtt 100%-ot ad ki.

Az output:

Korrelációs mátrixok (**A, B és C)**
Kanonikus korrelációk (sajátértékek négyzetgyöke):

```
> cancor(gen[,2:5],gen[,6:10])
$Summary
       R RSquared     LR ApproxF NumDF   DenDF pvalue
1 0.8793   0.7731 0.0795  1.3839    20 24.1662 0.2215
2 0.7463   0.5570 0.3506  0.8693    12 21.4575 0.5871
3 0.4116   0.1694 0.7914  0.3724     6 18.0000 0.8870
4 0.2173   0.0472 0.9528     NaN     2     NaN    NaN
```

Egyik kanonikus változó sem szignifikáns. Nincs bizonyítva a kapcsolat. Valószínűleg túl kicsi a minta.

Együtthatók:

```
> cancor(gen[,2:5],gen[,6:10])
$cor
[1] 0.8792722 0.7463372 0.4116297 0.2172688

$xcoef
              [,1]         [,2]          [,3]         [,4]
Alt  -1.022297e-05  0.000069722 -0.0003276579 0.0001429662
prec  1.143022e-02 -0.018091855 -0.0110937145 0.0158605207
max  -2.803969e-02  0.022719841 -0.0228538529 0.0646749640
min   1.130454e-03 -0.021851544 -0.0853475756 0.0174720672

$ycoef
              [,1]        [,2]       [,3]        [,4]
F0.40 -0.042105074  0.06781105 0.09541866  0.01773603
F0.60  0.031320794 -0.10452659 0.07057918 -0.08030188
F0.80  0.009003412 -0.05309862 0.05345934 -0.02076878
F1.00  0.018552591 -0.04425248 0.06325999 -0.02454376
F1.16  0.006436999 -0.07014658 0.08650252 -0.02715328
              [,5]
F0.40  0.002927490
F0.60  0.052022800
F0.80 -0.026731767
F1.00  0.003572987
F1.16  0.028507059
```

Az 1. kanonikus változó magas max. és min. hőmérséklettel, és alacsony magassággal és csapadék mennyiséggel korrelál.

```
$XUCorrelations
          U1      U2      U3      U4
Alt  -0.7663 -0.6245  0.1365 -0.0646
prec -0.8527  0.1545 -0.1484 -0.4764
max   0.8608  0.2796 -0.1423 -0.4008
min   0.7802  0.5606  0.1852  0.2067

$YVCorrelations
           V1      V2      V3      V4
F0.40  0.5680  0.4330 -0.2205  0.6566
F0.60  0.3870  0.1644  0.1205  0.8993
F0.80  0.7030 -0.2087  0.0690  0.4111
F1.00 -0.9222  0.2426 -0.1906 -0.2312
F1.16  0.3609 -0.4780 -0.0350 -0.7276
```

A kanonikus változó és az eredeti változók közötti korreláció. (Faktor struktúra). Négyzete méri az adott változó magyarázó hatását a kanonikus változóra nézve. 3 célra használjuk:

Interpretáció. Azon változókat, amelyeknek a korrelációja 0.3 felett van, tekintjük úgy hogy hozzájárulnak lényegesen a változóhoz.

A 2. csoport esetén az F1.00-val negatív a korreláció, a többivel pozitív. Így úgy tűnik, hogy a magas max. és min. hőmérséklet és alacsony magasság és csapadék mennyiség az F1.00 hiányával korrelál.