MTA SZTAKI

A kanonikus korrelációanalizis alkalmazása szivkoszoruér megbetegedések előrejelzésére

Lengyel Tamás

1. Bevezetés

A kanonikus korrelációanalizis alkalmazásának egy olyan lehetőségére szeretném ráirányitani a figyelmet, melynek segitségével bonyolult jelenségek időbeni lezajlására lehet következtetni.

Speciálisan szivkoszoruér megbetegedések előrejelzésére alkalmaztuk a módszert.

A vizsgálati adatokat az Országos Kardiológiai Intézettől kaptuk. Három községben, 5 évenként, összesen 3 alkalommal 1082 ember egészségi állapotára jellemző adatokat, dohányzási szokásokat regisztráltak.

Abból indultunk ki, hogy ezek az adatok sok információt tartalmazhatnak nemcsak az aktuális egészségi állapotról, hanem annak közeljövőben várható alakulásáról is. Ennek megfelelően a különböző alkalmakkor mért adatok közötti összefüggésre voltunk kivánesiak, valamint arra, hogyan lehet ezt az összefüggést felhasználni egyiküknek a másikkal történő becslése céljából. Az ilyen feladatokat a kanonikus korrelációanalizis segitségével lehet megoldani.

Ez a módszer az ilyen jellegü vizsgálatok körében ujszerü. Rövid idősorok esetében várható, hogy a szóban forgó módszer a vizsgálati adatok analizálásának hatékony segédeszköze lesz.

Az előadásban először ismertetem a kanonikus korrelációanalizis elvét, majd a konkrét példán bemutatom alkalmazását. Megjegyzem, hogy kevés értékelhető adat állt rendelkezésre, s ezért bizonyos egyszerüsitésekre kényszerültem. A módszer további finomitására is van lehetőség.

2. A kanonikus korrelációanalizis

A továbbiakban jelölje U_1 és U_2 azt a két valószínüségi változót (vektorváltozót), amelyek közötti kapcsolatot szeretnénk vizsgálni. Tegyük fel, hogy az összes komponensünk sztandardizált, azaz várható értékük nullával, szórásuk eggyel egyenlő. Jelölje cov (ξ, η) a ξ és η valószinüségi vektorváltozók kovariancia-mátrixát, azaz az (i,j)-edik eleme cov (ξ_i, η_j)-vel egyenlő, ahol ξ_i a ξ változó i. komponense. Legyen $\sum_{ij} = \text{cov}\,(U_i, U_j)$ $(i,j = 1,2)$. A feltételek miatt most a kovariancia- és a korrelációmátrix megegyezik. Ha az U_1 q-dimenziós, az U_2 (p-q)-dimenziós valószinüségi vektorváltozó /$p > q > 0$ egészek/, akkor tegyük fel, hogy rang $(\sum_{11}) = q$, rang $(\sum_{22}) = p-q$. Jelölje $m := \min\{q, p-q\}$, $k :=$ rang $(\sum_{12})$, A' illetve A^{-1} az A mátrix transzponáltját, illetve inverzét.

2.1. Két egydimenziós valószinüségi változó közötti összefüggés mérésének egy általánosan használt mérőszáma a korrelációs együttható. Ismeretes, hogy a két változó egymásra vonatkoztatott regressziós egyenesének az együtthatója és a korrelációs együttható között milyen kapcsolat van. Jelölje r az U_1 és U_2 korrelációs együtthatóját. Ekkor: $\hat{U}_2 = rU_1$ a regressziós egyenes, és $E(U_2 - \hat{U}_2)^2 = 1 - r^2$, amit ugy is mondhatunk, hogy az U_2 szórásnégyzetéből (vagy röviden U_2-ből) r^2-nyit magyaráz az U_1 változó.

2.2. Ugyancsak közismert mérőszám az egydimenziós és a többdimenziós valószinüségi változók közötti kapcsolat mérésére az un. többszörös korreláció. Ez a regresszióanalizisben éppen az U_2 egydimenziós változónak és az U_1 vektorváltozó lineáris függvényével való legkisebb négyzetes becslésének $(\hat{U}_2)$ a korrelációs együtthatója $(R_{U_2 \cdot U_1})$. A többszörös korreláció ugyanakkor a maximális korreláció az U_1 és U_2 lineáris függvényei között. Most

$$E(U_2 - \hat{U}_2)^2 = 1 - R^2_{U_2 \cdot U_1}.$$

Tehát U_2-ből $R^2_{U_2 \cdot U_1}$-nyit magyaráz az U_1 .

2.3. Általában is: az U_1 q-dimenziós és az U_2 p-q dimenziós valószinüségi vektorváltozók közötti kapcsolatot jellemezhetjük a lineáris függvényeik közötti maximális korrelációval. Ezt a számot nevezzük a két változó kanonikus korrelációjának. A fentiekből látható, hogy ez a korrelációs együttható fogalmának általánositása. Analitikus uton bevezethetünk $m = \min \{q, p-q\}$ darab kanonikus korrelációs együtthatót, illetve faktort. Ennek segitségével lehetőség van áttérni egy olyan koordinátarendszerre (vagy más szóval faktortérre), amelyben az U_1 és az U_2 komponensei korrelálatlanok, kivéve U_1 és U_2 ugyanolyan sorszámu koordinátáit, amelyek viszont "jól" korreláltak. Ebben a térben az U_1-nek az i. komponenséből az U_2 éppen annyit magyaráz, amennyit az i. koordinátája.

A két változó első kanonikus korrelációja alatt a következő értéket értjük:

$$\rho_1 := \max_{L_1 \in R^q,\ M_1 \in R^{p-q}} r(L_1'U_1, M_1'U_2),$$
$$D^2(L_1'U_1) = D^2(M_1'U_2) = 1$$

ahol $r(\xi, \eta)$ a ξ és η egydimenziós valószinüségi változók korrelációs együtthatóját, mig $D^2(\xi)$ a ξ szórásnégyzetét jelöli. Nyilván

$$D^2(L_1'U_1) = L_1'\sum_{11} L_1 = D^2(M_1'U_2) = M_1'\sum_{22} M_1 = 1.$$

Könnyen látható, hogy a fenti maximumot elérjük. $L_1'U_1$-et az első baloldali, $M_1'U_2$-öt az első jobboldali kanonikus faktornak nevezzük.

Az i. kanonikus korrelációt ($1 < i \leq m$), illetve (bal- és jobboldali) faktorokat a következőképpen definiáljuk:

$$\rho_i := \max_{L_i \in R^q,\ M_i \in R^{p-q}} r(L_i' U_1,\ M_i' U_2)$$

$$D^2(L_i' U_1) = D^2(M_i' U_2) = 1$$

$$\operatorname{cov}(L_i' U_1, L_j' U_1) = \operatorname{cov}(M_i' U_2, M_j' U_2) = 0, \quad 1 \leq j < i$$

A Lagrange multiplikátoros eljárással adódik, hogy:

$$\left(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \rho_i^2 \Sigma_{22}\right) M_i = 0 \qquad \text{és}$$

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \rho_i^2 \Sigma_{11} L_i = 0 \qquad (1 \leq i \leq m) \ .$$

Belátható, hogy $\rho_i = 0$, ha $k < i \leq m$ és

$$\operatorname{cov}(L_i' U_1, M_j' U_2) = \rho_i \delta_{ij} \qquad (i,j = 1,2,\ldots,m),$$

ahol δ_{ij} a Kronecker-delta.

Jelölje L illetve M az m darab baloldali, illetve jobboldali kanonikus együtthatóból, mint oszlopvektorból összeállított mátrixot, Λ az m kanonikus korrelációból összeállított diagonális mátrixot.

Tegyük fel, hogy $p = 2q$ és $k = m(=q)$. Nyilván $M'U_2$ komponenseinek az $\{L_1' U_1, L_2' U_1, \ldots, L_k' U_1\}$ baloldali kanonikus faktorokra vonatkozó legkisebb négyzetes becslésére (azaz regressziós síkjára) soronként értve a következő adódik:

$$\hat{M'U_2} = \Lambda\ L' U_1 \ . \qquad /1/$$

Könnyen belátható, hogy az U_2 eredeti komponenseinek legkisebb négyzetes becslésére az előbbi koordinátarendszerben (soronként értve):

$$\hat{U}_2 = M'^{-1}\ \Lambda\ L' U_1 \ . \qquad /2/$$

Tehát a kanonikus korrelációanalizis segitségével U_2-re egyszerüen adódik a legkisebb négyzetes regressziós közelités a fent emlitett faktortérben.

A 2.1. és a 2.2. pontban láttuk, hogy milyen szoros kapcsolat van a szórásmagyarázat és a változók összefüggését mérő korreláció között. A szórásmagyarázat általánositásával az összefüggés mérésének általánositásához juthatunk. Két valószinüségi vektorváltozó közötti kapcsolat nagyságának a mértékét azzal a két számmal jellemezhetjük, amely azt mutatja, hogy az egyik változó a másik összszórásnégyzetéből átlagosan mennyit magyaráz a legkisebb négyzetes becslések révén

$$(R_{U_2 \cdot U_1} \quad \text{és} \quad R_{U_1 \cdot U_2}) .$$

Attól függően, hogy az U_2-t és az U_1-et milyen koordinátarendszerben irjuk fel, ezek a számok különböző értéket vehetnek fel. Ezek után az U_2 és U_1 változók közötti összefüggést kétféleképpen jellemezhetjük:

a.) az $M'U_2$ és $L'U_1$ változók közötti összefüggéssel, azaz a kanonikus faktorok terében. Mivel itt /1/ szerint $\widehat{M'U_2} = \Lambda L'U_1$ és hasonlóan $\widehat{L'U_1} = \Lambda M'U_2$, igy:

$$\rho^2 = R_{U_2 \cdot U_1} = R_{U_1 \cdot U_2} = \frac{1}{k} \cdot \sum_{i=1}^{k} \rho_i^2$$

(Cramer mértéke: (1))

b.) az U_1 és $M'U_2$, illetve az U_2 és $L'U_1$ változók közötti összefüggéssel (Cooley-Lohnes-féle szórásmagyarázat: (2)). Ekkor /2/ szerint

$$\hat{U}_2 = M'^{-1} \Lambda L'U_1$$

és általában $$R_{U_2 \cdot U_1} \neq R_{U_1 \cdot U_2} .$$

3. A feladat konkrét megoldása

Jelölje U_1 egy vizsgált személy adataiból alkotott valószinüségi vektorváltozót és U_2 ugyanezen személy, ugyanezen változókra vonatkozó 5 év mulva mért adatait. Az a feladat, hogy az U_1 és U_2 közötti kapcsolat mértékét meghatározzuk és azt U_2-nek az U_1-gyel való becslésére felhasználjuk.

Kiválasztottuk a következő 6 változót: Broka index, systoles és diastoles vérnyomás, vitalkapacitás, serum cholesterin, body mass index. Az ismert egyéb adatok segitségével szétválasztottuk a végig egészséges (EEE, 209 db), az első vizsgálatnál már beteg (B, 28 db), illetve a második vizsgálatra megbetegedett (EB, 17 db) személyeket. Ellenőriztük, hogy a kiválasztott változók alapján is szétválaszthatók-e az egészséges és beteg személyek diszkriminancia analizis segitségével. Elsőfaju hibának (h_1) az egészségesek, másodfajunak (h_2) a betegek hibás klaszszifikációjának a relativ gyakoriságát nevezzük. Azt kaptuk, hogy $h_1 = 0,30$, $h_2 = 0,31$. Az EEE-re, az EB-re, valamint egyesitett halmazukra végrehajtottuk a kanonikus analizist (itt U_1 az első, U_2 a második vizsgálat adatait jelöli, segitségükkel becsüljük a $\sum_{ij}$ kovariancia-mátrixokat). Megállapitottuk, hogy az egészségesek beteggé válása során (EB) nagyobb az összefüggés U_1 és U_2 között, mint akkor, amikor végig egészségesek maradnak (EEE).

	/2/ szerint (b)		/1/ szerint (a)
	$R_{U_1 \cdot U_2}$	$R_{U_2 \cdot U_1}$	ϱ (Cramer)
EEE	0.574	0.570	0.727
EB	0.734	0.735	0.815
EEE és EB	0.579	0.573	0.730

Mindhárom esetben elkészitettük az $\hat{U}_2$ becsléseket a /2/ szerint. A diszkriminátor függvények segitségével osztályoztuk ezeket. Azt kaptuk, hogy a harmadik regresszió adja a legkisebb összhibát.

Regresszió /2/ szerint	hiba elsőfaju	másodfaju
EEE	0.25	0.35
EB	0.93	0.06
EEE és EB	0.24	0.35

Tehát eredményeink igazoljak, hogy ezzel a módszerrel nemcsak kimutatható az összefüggés az aktuális és a várható állapot között, hanem ez regresszió- és diszkriminancia analizis segitségével a következtetések levonására is felhasználható.

A módszert orvosi oldalról ugy lehetne finomitani, hogy további változók megválasztásával és mérésével a vizsgált jelenségek teljesebb leirását nyerjük. Azokat a változókat (pl.: a nem folyronosakat), amelyeket a kanonikus analizisbe nem tudjuk bevonni, más eszközök alkalmazásával a módszer finomitására ugyancsak felhasználhatjuk. Pl. a dohányzási szokás, az életkor esetében a különböző szokások, illetve korcsoportok rögzitése mellett hajtjuk végre az analizist. Igy ezen változók hatását is megvizsgálhatjuk.

Irodalom

(1) Anderberg, M.R.: Cluster Analysis for Applications, Academic Press, New York - London, 1973.

(2) Cooley, W.W., Lohnes, P.R.: Multivariate Data Analysis, John Wiley and Sons, New York, 1971.

(3) Rao, C.R.: Linear Statistical Inference and Its Applications, Wiley, New York, 1965.