Kanonična korelacijska analiza

Kanonična korelacijska analiza (Van de Geer, 1971; Cooley in Lohnes, 1971; Morrison, 1976) je multivariatna statistična metoda. Za dve dani množici spremenljivk nam kanonična korelacijska analiza omogoči poiskati taki linearni kombinaciji iz vsake množice spremenljivk, da je korelacija med njima maksimalna. Takih parov linearnih kombinacij (kanoničnih rešitev) je lahko več: prva kanonična rešitev poišče taki dve linearni kombinaciji, da bo korelacija med njima največja. Druga kanonična rešitev je naslednji par linearnih kombinacij spremenljivk z največjo korelacijo med njima, tako da sta ti dve spremenljivki neodvisni od kanoničnih spremenljivk prve rešitve. Postopek se podobno nadaljuje do več kanoničnih rešitev, vendar največ do manjšega števila spremenljivk v obeh množicah spremenljivk.

Oglejmo si kanonično korelacijsko analizo natančneje. Naj bo X matrika reda nXm_1, kjer je n število enot in m_1 število standardiziranih spremenljivk prve množice spremenljivk, in Y matrika reda nXm_2, kjer je m_2 število standardiziranih spremenljivk druge množice. Naj bo U linearna kombinacija x-spremenljivk; potem je U spremenljivka

$$U = X\,c \qquad (1)$$

kjer je c vektor uteži. Podobno je

$$V = Y\,d \qquad (2)$$

linearna kombinacija y-spremenljivk, kjer je d vektor uteži.

Potem lahko zapišemo naslednje korelacijske matrike

$$R_{xx} = X^{\backprime} X / n$$
$$R_{xy} = X^{\backprime} Y / n \quad , \quad R_{xy} = R_{yx}^{\backprime}$$
$$R_{yy} = Y^{\backprime} Y / n$$

kjer R_{xx} in R_{yy} ne smeta biti singularni matriki.

Naj bosta tudi vektorja U in V standardizirana. Tedaj velja

$$U^{\backprime} U / n = c^{\backprime} R_{xx} c = 1$$
$$V^{\backprime} V / n = d^{\backprime} R_{yy} d = 1$$

in

$$U^{\backprime} V / n = c^{\backprime} R_{xy} d = d^{\backprime} R_{yx} c = r$$

Poiščimo c in d tako, da bo r maksimalen! Pomagali si bomo z metodo Lagrangevih multiplikatorjev. Zato najprej definiramo funkcijo

$$F = c^{\backprime} R_{xy} d - \tfrac{1}{2}\mu(c^{\backprime} R_{xx} c - 1) - \tfrac{1}{2}\lambda(d^{\backprime} R_{yy} d - 1)$$

V ekstremnih točkah sta parcialna odvoda funkcije F po c in d enaka nič. Dobimo pogoja

$$R_{xy} d - \mu R_{xx} c = 0 \qquad (3)$$
$$R_{yx} c - \lambda R_{yy} d = 0 \qquad (4)$$

Iz obeh enačb sledi, da je $\mu = \lambda = r$. Iz druge enačbe lahko zapišemo

$$d = r^{-1} R_{yy}^{-1} R_{yx} c$$

Če to vstavimo v enačbo (3) in če na levi strani pomnožimo z R_{xx}^{-1}, dobimo

$$R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} c = r^2 c \qquad (5)$$

kjer predpostavimo, da obstojata inverzni matriki R_{xx}^{-1} in R_{yy}^{-1}.

Iz zadnje enačbe lahko c izračunamo kot lastne vektorje nesimetrične matrike, kjer so r^2 pripadajoče lastne vrednosti. Največji kanonični korelacijski koeficient določimo tako, da vzamemo tisti vektor c, ki ima največjo pripadajočo lastno vrednost.

Algoritmi za reševanje problema lastnih vrednosti običajno predpostavljajo simetričnost matrik. To dosežemo s substitucijo

$$c = R_{xx}^{-1/2} q$$

po kateri dobi enačba (5) obliko

$$R_{xx}^{-1/2} R_{xy} R_{yy}^{-1} R_{yx} R_{xx}^{-1/2} q = r^2 q$$

Ker je matrika leve strani enačbe reda $m_1 X m_1$, dobimo m_1 (,če je $m_1 \leq m_2$) lastnih vektorjev q in pripadajočih lastnih vrednosti. Sestavimo matriko Q, kjer so lastni vektorji stolpci. Matrika $Q`Q$ je potem diagonalna. Če lastne vektorje normaliziramo, je

$$Q` Q = I$$

Upoštevajmo zgoraj sestavljeno matriko Q in podobno posplošimo enačbo (1)

$$Q = R_{xx}^{1/2} C$$

in dalje

$$Q' Q = C' R_{xx} C = I$$

Po drugi strani je U=XC in zato

$$U' U / n = C' R_{xx} C$$

Torej uteži C določajo množico neodvisnih in standardiziranih linearnih kombinacij. Podobno velja za V=YD.

Pokazati se da, da je

$$U' V / n = C' R_{xy} D = P$$

kjer je P diagonalna matrika, v kateri so diagonalni elementi kanonični korelacijski koeficienti. To pomeni, da če vzamemo linearno kombinacijo iz U in linearno kombinacijo iz V, je korelacijski koeficient med njima različen od nič le, če imata isti indeks i; korelacijski koeficient je r_i.

Od vseh možnih kanoničnih rešitev so pomembne tiste, ki imajo statistično značilne kanonične korelacijske koeficiente, ki jih preverjamo z Wilksovo lambda statistiko. Ta je porazdeljena približno kot χ^2 porazdelitev z $(m_1-k)(m_2-k)$ prostostnimi stopnjami, kjer je k število že izločenih kanoničnih rešitev.