Analiza kanoniczna

Paweł Strawiński, Maciej Nasiński, Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

October 24, 2021

Plan zajęć I

Wprowadzenie

- Analiza kanoniczna jest uogólnieniem liniowej regresji wielorakiej na dwa zbiory zmiennych (np. zbioru zmiennych objaśniających X_i i zbioru zmiennych objaśnianych Y_i)
- W przeciwienstwie do regresji liniowej Y to najczesciej kilka zmiennych
- Technika polega na interpretacji zależności pomiędzy dwoma typami nowych zmiennych: zmiennymi kanonicznymi
- Pierwszy typ zmiennych kanonicznych jest liniową funkcją pierwszego zbioru zmiennych wejściowych, a drugi liniową funkcją drugiego zbioru zmiennych wejściowych
- Zmienne kanoniczne mają maksymalnie wyjaśniać zależności liniowe (korelacje) pomiędzy obydwoma zbiorami zmiennych

Pary zmiennych kanonicznych

- Cel to maksymalizacja kwadratu współczynnika korelacji między zmiennymi kanonicznymi
- Pierwsza para zmiennych kanonicznych wyjaśnia <u>większość</u> związków pomiędzy zbiorami zmiennych wejściowych, ale żeby w pełni opisać związki pomiędzy nimi potrzeba wyznaczyć kolejne pary zmiennych
- Żadna ze zmiennych należących do kolejnej pary zmiennych kanonicznych nie jest skorelowana z żadną ze zmiennych kanonicznych tego samego typu, gdyż wyjaśnia zależności między zbiorami zmiennych wejściowych w innych wymiarach
- Korelacje pomiędzy kolejnymi parami zmiennych kanonicznych są coraz słabsze

Przykłady pytań i problemów badawczych

- Jak **wyniki z egzaminów maturalnych** (jest ich kilka) studenta wpływają na jego **wyniki z przedmiotów**: mikroekonomia 1; algebra liniowa; podstawy prawa; język angielski?
- Jaki jest związek pomiędzy **charakterystykami respondenta** a jego **zadowoleniem**: z pracy, z życia osobistego, z sytuacji finansowej?
- Jaki jest związek pomiędzy **parametrami ciała** a **wynikami** poszczególnych **testów sprawnościowych**?

Założenia

- Dysponujemy dwoma **zbiorami** zmiennych: $Y_i = (Y_1, \ldots, Y_n)$ i $X_i = (X_1, \ldots, X_m)$
- Naszym zadaniem jest znalezienie takiej kombinacji liniowej zmiennych ze zbioru Y_i, która możliwie najsilniej koreluje z kombinacja liniowa zmiennych ze zbioru X_i
- Oznacza to, ze szukamy wektorów współczynników a_i i b_i takich, że korelacja $a_i' X_i$ i $b_i' Y_i$ jest możliwie największa

Tworzenie zmiennych kanonicznych

- **Zmienne kanoniczne** to kombinacje liniowe zbiorów zmiennych wejściowych:

$$U = A'X_i$$

$$V = B'Y_i$$

- $U = [u_{li}]$ to macierz zmiennych kanonicznych pierwszego typu; u_{li} to wartość l-tej zmiennej w i-tym obiekcie - scores
- $V = [v_{li}]$ to macierz zmiennych kanonicznych drugiego typu; v_{li} to wartość l-tej zmiennej w i-tym obiekcie - scores
- $A' = [a_{jl}]$ to transponowana macierz wag kanonicznych, a_{lj} to waga kanoniczna j-tej zmiennej w zbiorze X_i dla l-tej zmiennej kanonicznej pierwszego typu - coef/rotations
- $B' = [b_{jl}]$ to transponowana macierz wag kanonicznych, b_{lj} to waga kanoniczna j-tej zmiennej w zbiorze Y_i dla l-tej zmiennej kanonicznej drugiego typu

Tworzenie zmiennych kanonicznych: założenia

- Wektory u_i i u_j są nieskorelowane między sobą
- Wektory v_i i v_j są nieskorelowane między sobą
- Korelacje $corr(u_i; v_i)$ tworzą nierosnący ciąg odpowiadający możliwie największym cząstkowym korelacjom

Obliczanie wag kanonicznych 1

- **Wagi kanoniczne** mają maksymalizować korelację pomiędzy kolejnymi parami zmiennych kanonicznych: **korelację kanoniczną**
- Wyznacza się je w oparciu o łączną macierz korelacji zmiennych:

$$R = \begin{bmatrix} R_{YY} & R_{XY} \\ R_{YX} & R_{XX} \end{bmatrix}$$

- R_{YY} to macierz korelacji zmiennych objaśnianych Y_i
- R_{XX} to macierz korelacji zmiennych objaśniających X_i
- R_{XY}, R_{YX} to macierze korelacji obu rodzajów zmiennych

Obliczanie wag kanonicznych 2

- Na początku poszukujemy wag kanonicznych pierwszej pary zmiennych kanonicznych, ponieważ ta para w największym stopniu wyjaśnia zależności pomiędzy zbiorami X_i i Y_i; następnie wag kanonicznych kolejnych par zmiennych kanonicznych
- Wagi kanoniczne maksymalizują współczynnik korelacji kanonicznej:

$$r_{u_i,v_i} = \frac{(a_i' R_{XY} b_i)}{\sqrt{(a_i' R_{XX} a_i)(b_i' R_{YY} b_i)}}$$

Obliczanie wag kanonicznych 3

- Wagi kanoniczne wyznacza się poprzez rozwiązanie układów równań jednorodnych o postaci:

$$\begin{cases} (R_{XX}^{-1}R'_{XY}R_{YY}^{-1}R_{YX} - \lambda_i I)a_i = 0 \\ \\ (R_{YY}^{-1}R'_{YX}R_{XX}^{-1}R_{XY} - \lambda_i I)b_i = 0 \end{cases}$$

- λ_i to pierwiastek charakterystyczny (wartość własna) odpowiedniej macierzy

Obliczanie wag kanonicznych 4

- Liczba niezerowych pierwiastków charakterystycznych równań wyznacznikowych jest równa $s = min(n, m)$
- Po malejącym uporządkowaniu wartości własnych znajdujemy wagi kanoniczne dla kolejnych par zmiennych kanonicznych, wstawiając do układu równań kolejne wartości własne

Rozwiązanie

- Wagi kanoniczne określają wkład poszczególnych zmiennych wejściowych w tworzenie zmiennych kanonicznych
- Zmienne kanoniczne danego typu nie są ze sobą skorelowane, dlatego suma kwadratów współczynników korelacji kanonicznej dla wszystkich par zmiennych kanonicznych stanowi miarę stopnia **wyjaśnienia zmienności poprzez związki liniowe zbioru zmiennych objaśnianych przez zbiór zmiennych objaśniających**

$$R^2 = \sum_{l=1}^{s} r^2_{u_i,v_i}$$

Ładunki

- Żeby zinterpretować zmienne kanoniczne przedstawiamy zbiory zmiennych wejściowych jako kombinacje liniowe zmiennych kanonicznych:

$$Y_i = CV$$

$$X_i = DU$$

- $C = [c_{jl}]$ to macierz kanonicznych ładunków czynnikowych, c_{jl} jest kanonicznym - loadings ładunkiem czynnikowym znajdującym się przy j-tej zmiennej wejściowej i l-tej zmiennej kanonicznej pierwszego typu
- $D = [d_{il}]$ to macierz kanonicznych ładunków czynnikowych, d_{il} jest kanonicznym ładunkiem czynnikowym znajdującym się przy i-tej zmiennej wejściowej i l-tej zmiennej kanonicznej drugiego typu - loadings

Ładunki – cd

- Kanoniczne ładunki czynnikowe są współczynnikami korelacji liniowej pomiędzy zmiennymi pierwotnymi a zmiennymi kanonicznymi

$$c_{jl} = r_{y_j, v_l}; j = 1, 2, \ldots, n; l = 1, 2, \ldots, s;$$

$$d_{il} = r_{x_i, v_l}; i = q + 1, q + 2, \ldots, n + m; l = 1, 2, \ldots, s;$$

- Im większa wartosc bezwzględna ładunku czynnikowego, tym większy nacisk należy kłaść na daną zmienną przy interpretacji zmiennej kanonicznej
- Przy interpretacji zmiennych kanonicznych bierzemy pod uwagę zmienne wejściowe silnie skorelowane

Uwagi praktyczne

- Analizowane zmienne powinny mieć rozkład wielowymiarowy normalny
- W zbiorze danych nie występują obserwacje odstające (miara Cooka, wartości dźwigni, etc.)
- Zmienne nie są również współliniowe
- Próba musi mieć dostatecznie dużą liczebność (nieformalnie: liczba obserwacji powinna być większa od co najmniej 20$\times$liczba zmiennych)

Określanie liczby par zmiennych kanonicznych – założenia

- Zakładamy, że co najmniej k pierwszych korelacji kanonicznych jest istotnych i testujemy hipotezę o istotności ostatnich $s - k$ korelacji kanonicznych
- Wykorzystujemy statystykę testową χ^2
- Weryfikacja istotności par zmiennych kanonicznych odbywa się w sposób iteracyjny
- Jeśli wartość statystyki testowej jest mniejsza od wartości krytycznej przy przyjętym poziomie istotności, to przynajmniej jeden współczynnik korelacji kanonicznej o indeksie $k + 1$ jest istotny

Określanie liczby par zmiennych kanonicznych – cd

- Wiedząc, że kolejne korelacje kanoniczne są coraz mniejsze, przyjmujemy na początek procesu weryfikacji $k = 0$ – co najmniej pierwsza z korelacji kanonicznych jest istotna
- Po braku podstaw do odrzucenia hipotezy o istotności, zwiększamy kolejno indeks k o jeden i testujemy istotność kolejnych współczynników korelacji kanonicznej
- W ostatecznej analizie uwzględniamy wszystkie pary zmiennych kanonicznych, dla których współczynniki korelacji kanonicznej są istotne

Wariancja wyodrębniona

- Dzieląc sumy kwadratów współczynników korelacji danej zmiennej kanonicznej przez liczbę zmiennych wejściowych odpowiedniego typu uzyskujemy wartość **wariancji wyodrębnionej**
- Wartość ta określa jaki procent wariancji zmiennych wejściowych wyjaśnia średnio dana zmienna kanoniczna

$$\bar{R}^2_{u_l} = \frac{1}{n}\sum_{j=1}^{n} c^2_{jl}, l = 1, 2, \ldots, s;$$

$$\bar{R}^2_{v_l} = \frac{1}{m}\sum_{j=n+1}^{n+m} d^2_{jl}, l = 1, 2, \ldots, s;$$

Współczynniki redundancji

- **Współczynniki redundancji** są miarą stopnia wyjaśnienia wariancji zmiennych pierwotnych danego typu przez zmienne kanoniczne drugiego typu:

$$R^2_{v_l, Y_i} = \bar{R}^2_{v_l} \lambda_l, l = 1, 2, \ldots, s;$$

$$R^2_{u_l, X_i} = \bar{R}^2_{u_l} \lambda_l, l = 1, 2, \ldots, s;$$

- Czyli dowiadujemy się, na ile nadwymiarowy jest jeden zbiór danych wobec drugiego zbioru danych