Analiza Wielowymiarowa

Metody czynnikowe

Maciej Nasiński, Paweł Strawiński, Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 6
30 listopada 2023

Wprowadzenie

- Metody czynnikowe stanowią zbiór metod i procedur statystycznych pozwalających na redukcję dużej liczby zmiennych do kilku wzajemnie nieskorelowanych czynników
- Za ich pomocą można zachować stosunkowo dużą część informacji zawartych w zmiennych pierwotnych
- Jednocześnie każda z tych metod niesie inne treści merytoryczne

Cele

- **Redukcja liczby zmiennych** bez istotnej utraty zawartych w nich informacji
- **Transformacja** układu zmiennych w nowy układ czynników głównych
- **Tworzenie skal i miar** na podstawie wartości kilku zmiennych
- **Ustalanie wag** określających znaczenie, jakie należy przypisać poszczególnym zmiennym podczas analizy
- **Ortogonalizacja przestrzeni**, w której rozpatrywane są analizowane obiekty
- **Wykrywanie ukrytych związków** między zmiennymi
- **Opis zjawisk** za pomocą nowych kategorii zdefiniowanych przez czynniki

Przykłady zastosowań

- Kiedy interesuje nas **eksploracja** i rozpoznanie struktury zbioru danych
- Gdy nie posiadamy modelu „głębokiej” **struktury czynników** wyjaśniających związki między danymi
- Gdy potrzebujemy **zredukować zbiór zmiennych** skorelowanych ze sobą do wykorzystania ich w postaci zagregowanej w późniejszych etapach analizy
- Gdy chcemy stworzyć **skalę, indeks, miernik** ukrytego zjawiska i jednoznacznie obliczyć jego wartość

Przykładowe pytania i zagadnienia badawcze

- Stworzenie indeksu kapitału społecznego (FA)
- Wypowiedzenie się na temat postawy respondentów w oparciu o wiele stwierdzeń dotyczących jednego zagadnienia (np. zadowolenia ze spędzania czasu wolnego) (FA lub PCA)
- Stworzenie agregatowej zmiennej z wartości pomiarów potrzebnej do dalszej analizy (PCA)
- Stworzenie zmiennej opisującej objawy depresji, do wykorzystania w regresji liniowej, celem uniknięcia silnego skorelowania zmiennych (PCA)

Dwa modele metod czynnikowych

- Model **klasyczny**, w którym wariancję całkowitą zmiennych dzieli się na wariancję wspólną i wariancję specyficzną (klasyczna analiza czynnikowa)
- Model **komponentowy**, w którym nie uwzględnia się struktury wariancji (metoda składowych głównych)

Ogólna charakterystyka metody

- Klasyczna analiza czynnikowa (ang. *factor analysis* (FA)) służy do znajdowania ukrytych czynników, które określają związki pomiędzy zmiennymi pierwotnymi (obserwowanymi)
- Zbiór zmiennych pierwotnych dzieli się na podzbiory, które są silnie determinowane przez określoną grupę czynników (ukrytych) a słabiej przez pozostałe
- Jest to metoda modelowania liniowego – zakłada się, że zmienne można przedstawić za pomocą liniowej funkcji zmiennych ukrytych (czynników)
- Nie ma podziału na zmienne objaśniające i objaśniane

Obszar zastosowania analizy czynnikowej

1. Analiza wyjaśniająca (eksploracyjna)
   - Czynniki są opisywane przez grupowanie w zbiory zmiennych najsilniej ze sobą skorelowanych;
   - Technika ma za zadanie wykryć zależności pomiędzy zmiennymi pierwotnymi, a zmiennymi ukrytymi bez wstępnych założeń dotyczących kierunku tych powiązań;
2. Analiza potwierdzająca (konfirmacyjna)
   - Za jej pomocą weryfikujemy poprawność teorii, czyli hipotez badawczych o występowaniu pewnych nieobserwowanych zjawisk
   - Technika ta weryfikuje określoną strukturę czynników, w której zmienne pierwotne zależą od domniemanych lub znanych badaczowi zmiennych ukrytych

Zależność funkcyjna

$$X_i = f(F_1, F_2, \ldots, F_p) + \varepsilon_i$$

- p – liczba zmiennych ukrytych
- k – liczba zmiennych pierwotnych
- X_i – zmienna wyjściowa, o której zakłada się, że ma rozkład normalny $(i = 1, ..., k)$
- F_j – zmienna ukryta, czynnik $j = 1, 2, ..., p$, gdzie $p \leq k$
- ε_i – czynnik losowy, zakłada się, że jest to zmienna losowa o rozkładzie normalnym

Zależność funkcyjna cd.

$$X_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{ip} f_p + \varepsilon_i$$

$$X = \Lambda f + \varepsilon$$

- λ_j – waga stojąca przy j-tej zmiennej ukrytej dla i-tej zmiennej pierwotnej, inaczej ładunek czynnikowy

Założenia dotyczące wariancji

$$\sigma_i = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ik}^2 + \varphi = \sum_{j=1}^{k} \lambda_{ij}^2 + \varphi$$

$$Cov(X_i, X_j) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{ip}\lambda_{jp}$$

$$\sum = \Lambda\Lambda' + \Phi$$

- σ_i – wariancja zmiennej wyjściowej X_i
- $\sum_{j=1}^{p} \lambda_{ij}^2$ – zmienność wspólna zmiennej wyjściowej X_i
- φ – zmienność swoista zmiennej wyjściowej X_i
- $\sum$ – macierz kowariancji, Φ to macierz z wartościami swoistymi na przekątnej (szacowana za pomocą macierzy kowariancji lub korelacji z próby)

Założenia dodatkowe

- Czynniki wspólne nie są skorelowane ze sobą
- Czynniki swoiste (inaczej specyficzne) nie są skorelowane ze sobą
- Czynniki wspólne i swoiste nie są ze sobą skorelowane
- Wartości czynników wspólnych są wystandaryzowane: $E(F_j) = 0$ i $Var(F_j) = 1$

Algorytm

- Szukamy oszacowań ładunków dla czynników oraz dla części wspólnej wariancji
- Po określeniu rozwiązania początkowego w następnym kroku można dokonać rotacji czynników w celu łatwiejszej interpretacji
- Rozwiązujemy względem $\hat{\Lambda}$ i $\hat{\Phi}$ ograniczenie:

$$S = \hat{\Lambda}\hat{\Lambda}' + \hat{\Phi}$$

Metoda osi głównych

- Metodę osi głównych stosuje się przy wyznaczaniu współczynników głównych składowych
- Jedyna różnica, w stosunku do procedury stosowanej w analizie głównych składowych, polega na wykorzystaniu zredukowanej macierzy korelacji zamiast pełnej macierzy korelacji
- Na głównej przekątnej zredukowanej macierzy korelacji zamiast jedynek znajdują się wartości zasobów zmienności wspólnej kolejnych zmiennych pierwotnych

Metoda centroidalna

- Opiera się na geometrycznym podejściu do analizy czynnikowej
- Kolumny macierzy danych wejściowych można interpretować jako konfigurację m wektorów zmiennych w n wymiarowej przestrzeni euklidesowej R_n. Wzajemny układ wektorów, reprezentujących zmienne, określa korelacje pomiędzy zmiennymi, tzn. cosinusy kątów między wektorami są równe współczynnikom korelacji pomiędzy zmiennymi
- Zakłada się, że osie poszczególnych czynników przechodzą przez środki ciężkości (centroidy) konfiguracji wektorów
- Kolejne czynniki wyjaśniają maksymalną część zmienności wspólnej zmiennych pierwotnych
- Wartości ładunków czynnikowych to współrzędne punktów reprezentujących zmienne w nowym, ortogonalnym układzie odniesienia

Metoda największej wiarogodności

- W przeciwieństwie do innych metod, należy określić liczbę czynników wspólnych, którą chcemy uzyskać **przed przystąpieniem do analizy**
- Założenie: dane wejściowe, zmienne wyjściowe, składnik losowy oraz funkcje zmiennych ukrytych pochodzą z próby o wielowymiarowym rozkładzie normalnym
- Postać analityczna funkcji wiarogodności:

$$L = -\frac{1}{2} n\{ln|\Lambda\Lambda' + \Phi| + tr(S|\Lambda\Lambda' + \Phi|^{-1})\}$$

Rotacja czynników

- Uzyskana macierz ładunków czynnikowych nie jest jedynym możliwym rozwiązaniem analizy czynnikowej
- Można stworzyć nieskończenie wiele różnych macierzy ładunków czynnikowych poprzez obrót układu wzajemnie ortogonalnych osi
- Rotacja ma pomóc w znalezieniu układu, który będzie prostszy w interpretacji
- Istnieją dwie grupy metod rotacji: ortogonalne i ukośne

Rotacje ortogonalne

- Polegają na znalezieniu ortogonalnej macierzy transformacji
- Najbardziej znane metody to **varimax** i **quartimax**
- Varimax minimalizuje liczbę zmiennych potrzebnych do wyjaśnienia danego czynnika
- Quartimax minimalizuje liczbę czynników potrzebnych do wyjaśnienia danej zmiennej

Rotacje ukośne

- Macierz ładunków czynnikowych staje się macierzą wzorców zachowań
- Do wyznaczenia korelacji czynników wykorzystuje się wagi nadane poszczególnym czynnikom F

Metody wyboru liczby czynników

Do wyboru optymalnej liczby czynników można stosować następujące metody:

- Metodę wartości własnych większych od jedności
- Metodę procentu wariancji tłumaczonej przez czynniki
- Metodę testu osypiska

Ale i tak ostateczna decyzja jest subiektywnym wyborem badacza

Metoda wartości własnej większej od jedności

- Jest to najczęściej spotykana metoda: każdy czynnik powinien wyjaśniać zmienność co najmniej jednej zmiennej pierwotnej
- Polecana, jeżeli liczba zmiennych jest większa niż 20
- W przypadku analiz na mniejszych zbiorach danych, metoda ta ma tendencję do wybierania zbyt małej liczby czynników

Metoda procentu wariancji tłumaczonej

- Liczba wybranych czynników ustalana jest na podstawie procentu wariancji przez nie tłumaczonej
- Dążymy do odtworzenia co najmniej 70% wariancji (niższe wartości w przypadku dużych zbiorów danych)
- Żaden kolejny czynnik poza wybranymi nie tłumaczy więcej niż 5% wariancji

Metoda testu osypiska

- Najpierw sporządzamy wykres, na którym na osi poziomej umieszczamy kolejne czynniki, natomiast na osi pionowej ich wartości własne
- Szukamy punktów załamania, w których zmienia się kąt załamania krzywej (zaczynają się kolejne rumowiska)
- Miejsce punktu załamania określa maksymalną liczbę czynników kwalifikujących się do dalszej analizy
- Metoda ta pozwala włączyć do analizy większą liczbę czynników niż metoda wartości własnych większych od 1

Nazwy czynników

- Dla każdego czynnika wybieramy kilka zmiennych o najwyższych ładunkach czynnikowych
- Następnie probujemy nadać nazwę czynnikowi, wykorzystując najważniejsze zmienne
- Po ustaleniu nazw czynników należy podjąć próbę znalezienia odniesienie zmiennych do danego, głębszego wymiaru, ich związek z badaną zmienną ukrytą

Wskaźnik Kaisera-Mayera-Olkina (KMO)

- Informuje, czy istnieją podstawy do stosowania analizy czynnikowej
- Indeks o wartościach [0,1] porównuje cząstkowe współczynniki korelacji z dwuzmiennowymi współczynnikami korelacji

$$KMO = \frac{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2}{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2 + \sum_{i \neq j} \sum_{j \neq i} a_{ij}^2}$$

- r_{ij} – element macierzy korelacji R
- a_{ij} – współczynnik korelacji cząstkowej
- Im wartość wskaźnika bliższa 1, tym silniejsze podstawy do zastosowania analizy czynnikowej

Test Bartletta o sferyczności

- H_0: $R = I$ (macierz korelacji jest macierzą jednostkową)
- H_1: $R \neq I$ (macierz korelacji nie jest macierzą jednostkową)
- Celem jest odrzucenie H_0

Analiza składowych głównych (PCA)

- Stanowi metodę transformacji zmiennych pierwotnych we wzajemnie ortogonalne nowe zmienne, tzw. składowe główne
- Służy redukcji wymiaru przestrzeni cech oraz pogrupowaniu ich w podzbiory
- Dzięki niej można graficznie zaprezentować konfigurację porównywanych zmiennych

Ogólna charakterystyka – cd.

- Zmienne pierwotne poddaje się standaryzacji, więc ich wariancje są sobie równe
- Nowa agregatowa zmienna powinna wyjaśniać maksymalną część wariancji zmiennych pierwotnych
- Wariancja nowej zmiennej agregatowej jest nazywana wartością własną (ang. *eigenvalue*)
- Zbiór danych powinien być jednorodny (brak obserwacji odstających)

Zapis formalny modelu

$$PC_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ik}X_k$$

$$\sum_{j=1}^{k} w_{ij}^2 = 1$$

- Współczynniki w przy zmiennych X stanowią wagi, jakie przypisuje się zmiennym w tworzeniu składowej głównej
- Zakładamy, że poszukiwane czynniki są niezależne i mają wystandaryzowany rozkład normalny

Wyznaczanie współczynników

- Wartości wektora w są tak dobierane, aby maksymalizować wariancję PC
- Szukamy wartości własnych następującego układu równań:

$$|R - \lambda I| = 0$$

- R – macierz korelacji k zmiennych wyjściowych (pierwotnych)
- Λ – macierz wektorów własnych o wymiarach kxk
- Wariancja i-tej składowej to i-ta wartość własna

Wyznaczanie współczynników – cd

- Każdej wartości własnej odpowiada wektor własny macierzy o postaci:

$$Rw_i = \lambda_i w_i$$

- w_i – wektor własny macierzy korelacji
- Wartości składowe tego wektora stanowią wartości współczynników przy zmiennych pierwotnych; ich kombinacja tworzy nowe zmienne: składowe główne
- Pułapka: utworzona kombinacja liniowa jest zależna od jednostek miary i rzędów wielkości poszczególnych zmiennych (należy standaryzować zmienne!)

Wybór optymalnej liczby składowych głównych

- Dążymy do odtworzenia maksymalnej ilości informacji z pierwotnego zbioru zmiennych
- W praktyce wybieramy liczbę składowych, które łącznie wyjaśniają powyżej 70% zmienności zmiennych pierwotnych
- Nie uwzględnia się tych składowych głównych, dla których wartości własne są niższe od średniej
- Można opuścić składowe główne, dla których wartości własne są niższe od 1 (symulacje wskazują, że lepszym progiem jest 0,7)
- Opuszczamy składowe główne, które mają mniejszy udział w wariancji niż 5%

Wybór właściwego modelu

- Wybór między PCA a FA zależy przede wszystkim od celu analizy
- W klasycznej analizie czynnikowej mała liczba czynników pozwala wyjaśniać zależności pomiędzy zmiennymi obserwowanymi; chcemy zidentyfikować zmienne ukryte
- W analizie składowych głównych dążymy do zachowania jak największej ilości informacji przy jak najmniejszej liczbie nowych zmiennych; chcemy uprościć strukture danych
- FA to analiza modelowa, PCA to technika eksploracyjna, pomocnicza

FA i PCA – różnice

- **Wariancja** – FA obejmuje pewną część wariancji, zwaną wariancją wspólną PCA obejmuje wariancję całkowitą zmiennych
- **Punkt wyjścia** – dla FA to zredukowana macierz korelacji; dla PCA zwykła macierz korelacji
- **Zmienne pierwotne** – w FA zmienna pierwotna jest funkcją czynników wspólnych i swoistych; w PCA główna składowa jest funkcją zmiennych pierwotnych
- **Zależności** – w FA czynniki mogą być skorelowane; w PCA składowe główne są zawsze niezaleĹĽne