Wielowymiarowa Analiza Korespondencji

Arkadiusz Urbaniak (276034)

30 czerwca 2025

Spis treści

1 Wstęp

1.1 Wprowadzenie

Wielowymiarowa analiza korespondencji (ang. Multiple Correspondence Analysis, MCA) to metoda statystyczna służąca do eksploracyjnej analizy danych kategorycznych, będąca naturalnym rozszerzeniem klasycznej analizy korespondencji (Correspondence Analysis, CA). Jej głównym celem jest identyfikacja i wizualizacja związków między kategoriami wielu zmiennych jakościowych poprzez odwzorowanie ich w przestrzeni o zredukowanej liczbie wymiarów. Dzięki temu możliwe jest wychwycenie struktur i powiązań w zbiorach danych, w których klasyczne metody analizy wariancji czy regresji są nieadekwatne z powodu braku zmiennych ilościowych.

1.2 Motywacja do wyboru tematu

Zastosowanie MCA jest szczególnie użyteczne w analizie danych ankietowych, gdzie dominują odpowiedzi kategoryczne (np. płeć, wykształcenie, poglądy polityczne, postawy wobec zjawisk społecznych, preferencje konsumenckie). Zamiast analizować każdą zmienną osobno lub w parach, MCA pozwala spojrzeć na całość zbioru odpowiedzi i wychwycić wzorce. Z tego powodu zdecydowałem się podjąć analizy MCA, uważając ten rodzaj narzędzia za bardzo istotne, ale także naturalnie rozwijające już zdobytą wiedzę.

1.3 Cel raportu

Niniejsza praca przedstawia analizę MCA zarówno od strony teoretycznej, jak i praktycznej. Omówione zostały jej podstawowe założenia, sposób kodowania danych, przebieg obliczeń oraz metody wizualizacji i interpretacji wyników. Część praktyczna opiera się na danych symulowanych, które odwzorowują typową strukturę odpowiedzi w badaniu ankietowym. Celem pracy jest pokazanie, w jaki sposób MCA może zostać wykorzystana do identyfikacji wzorców

zachowań i postaw w danych jakościowych oraz jakie są ograniczenia tej metody w praktyce badawczej.

2 Wielowymiarowa Analiza Korespondencji

2.1 Podstawowe założenia i cele

Aby stosować wielowymiarową analizę korespondencji, zmienne, dla których będziemy przeprowadzać badania, powinny być zmiennymi kategorycznymi (jakościowymi), czyli zmiennymi, które mogą przyjmować jedną z ograniczonych możliwych wielkości. Oznacza to także, że nie operujemy bezpośrednio na wartościach liczbowych na podstawie których moglibyśmy wyliczyć np. wariancję, a zamiast tego korzystamy z częstości.

Kolejnym bardzo ważnym założeniem jest brak rozróżnienia między zmiennymi zależnymi i niezależnymi. Wszystkie zmienne traktowane są jako równorzędne i pełnią funkcję opisową, a nie przyczynową. Stanowi to znaczącą różnicę względem innych metod takich jak regresja, analiza wariancji czy modele klasyfikacyjne. Z analizy MCA nie można stwierdzić wynikania. Zamiast tego celem jest identyfikacja układów kategorii, które często współwystępują. Z tego powodu na podstawie analizy korespondencyjnej możemy stwierdzić zależność między zmiennymi ale nie możemy określić jej kierunku.

Analiza korespondencji nie radzi sobie również z brakami danych. Oznacza to, że przystępując do badania zależności między zmiennymi kategorycznymi musimy pierw pozbyć się w jakiś sposób danych zawierających niepełne informacje.

2.2 Kodowanie zmiennych jakościowych

2.2.1 Macierz kodów

Aby dane potrzebne do przeprowadzenia wielowymiarowej analizy korespondencji nadawały się do użycia, można zastosować tzw. Macierz kodów. Jest to macierz, w której dla każdej badanej jednostki odpowiadającej jednemu

wierszowi macierzy przyporządkowujemy w kolejnych kolumnach wartość 0 lub 1. Kolumny natomiast dzielimy ze względu na cechy, a następnie wewnątrz tej cechy możemy tylko dla jednej kategorii przypisać wartość 1, a dla pozostałych wstawiamy zera.

Tabela 1: Przykład macierzy kodów dla wielowymiarowego przypadku

Respondent	**Wiek**		**Płeć**		**Stanowisko**	
	Młody	Starszy	Kobieta	Mężczyzna	Pracownik	Kierownik
1	1	0	1	0	1	0
2	0	1	0	1	0	1
3	1	0	0	1	1	0
...	...	...	...	...	...	...
n	1	0	0	1	1	0

Taka forma reprezentowania danych jest prosta do uzyskania. Dodatkowo łatwo rozszerzyć macierz o dodatkowe cechy. Jest to istotne zwłaszcza w kontekście wielowymiarowej analizy korespondencji.

2.2.2 Tablica Burt'a

Macierz kodów ma jednak również swoje wady. Jedną z nich jest to, że dla dużej ilości obserwacji macierz osiąga ogromne rozmiary. Z tego powodu często korzysta się z Tablicy Burt'a. Oparta jest ona na iloczynie wewnętrznym macierzy kodów. Jeżeli dysponujemy tabelami częstości, to tablicę Burt'a otrzymamy dodając bloki określające związki każdej zmiennej z sobą samą.

Tabela 2: Przykład tablicy Burt'a dla wielowymiarowego przypadku

	Wiek		**Płeć**		**Stanowisko**	
	Młody	Starszy	Kobieta	Mężczyzna	Pracownik	Kierownik
Wiek: Młody	200	0	110	90	190	10
Wiek: Starszy	0	150	60	90	100	50
Płeć: Kobieta	110	60	170	0	130	40
Płeć: Mężczyzna	90	90	0	180	160	20
Stanowisko: Pracownik	190	100	130	160	290	0
Stanowisko: Kierownik	10	50	40	20	0	60

Aby uzyskać taką macierz, posiadając już macierz kodów nazwaną jako A, należy wykonać następujące działanie: $A^T A$. Wynikiem takiej operacji będzie wyżej otrzymana macierz Burt'a.

2.3 Metodyka przeprowadzania analizy

Załóżmy, że posiadamy już wcześniej przedstawioną macierz kodów. Skupmy się zatem na kolejnych krokach działań potrzebnych do przeprowadzenia MCA. Warto na początku zauważyć, że wykonywane operacje są w dużej mierze analogicznie do dwuwymiarowej analizy korespondencji.

2.3.1 Obliczenie macierzy częstości względnych

Zaczynamy od obliczenia macierzy częstości względnych daną wzorem

$$P = \frac{1}{n} A$$

gdzie n to liczba obserwacji.

2.3.2 Obliczenie mas wierszy i kolumn

Następnym krokiem jest obliczenie mas wierszy i kolumn.

- **Masa wiersza:**

$$r_i = \sum_j p_{ij}$$

- **Masa kolumny:**

$$c_j = \sum_i p_{ij}$$

2.3.3 Tworzenie macierzy reszt standaryzowanych

Kolejnym etapem jest stworzenie macierzy reszt standaryzowanych, której elementy dane są wzorem:

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

Z takiej macierzy odczytać można odchylenia od wartości oczekiwanych przy założeniu niezależności wierszy i kolumn.

2.3.4 Dekompozycja macierzy reszt standaryzowanych

Kolejno wykonujemy dekompozycję według wartości osobliwych (Singular Value Decomposition):

$$S = U \Sigma V^T$$

gdzie:

- U - macierz lewych wektorów osobliwych (wiersze)
- Σ - macierz diagonalna wartości osobliwych
- V - macierz prawych wektorów osobliwych (kolumny)

2.3.5 Wyznaczenie współrzędnych

Współrzędne wierszy wyznaczamy jako:

$$F = D_r^{-1/2} V \Sigma$$

Współrzędne kolumn wyznaczamy jako:

$$G = D_c^{-1/2} U \Sigma$$

gdzie D_r i D_c to macierze diagonalne mas wierszy i kolumn. Otrzymane macierze w kolejnych kolumnach zawierają współrzędne dla odpowiadających im wymiarów. Podczas analogicznego wyliczenia współrzędnych co w wyżej przedstawionej metodzie, otrzymujemy współrzędne do wykresów typu principal. Macierz F odpowiada współrzędnym dla wierszy, a G dla kolumn.

2.3.6 Przeprowadzanie MCA w Python

W środowisku Python znajduje się biblioteka "prince" obsługująca wielowymiarową analizę korespondencji. Jednakże na potrzeby tej pracy zdecydowałem się na samodzielne zaimplementowanie MCA. Wszystkie wykonane operacje podczas programowania odpowiadają przedstawionemu schematowi przeprowadzania analizy.

2.4 Interpretacja

Standardowo, ze względu na łatwość interpretacji, przyjęło się korzystanie z pierwszych dwóch znaczących wymiarów. W ten sposób możemy zaprezentować otrzymane punkty na kartezjańskim układzie współrzędnych. Z uzyskanych wykresów analizie poddaje się odległości między punktami, które dostarczają informacji o podobieństwie częstości względnych, jakie dane wiersze mają w odpowiednich kolumnach. Powszechną praktyką jest przedstawianie wierszy oraz kolumn na jednym wykresie. Należy jednak pamiętać, że w tym wypadku interpretacji poddawać można tylko odległości między punktami reprezentującymi wiersze, albo odległości między punktami reprezentującymi

kolumny. Oznacza to, że nie powinniśmy interpretować odległości między kolumnami a wierszami. Warto jednak zauważyć, że można formułować pewne ogólne spostrzeżenia na temat charakteru wymiarów. Opierać można się, dla przykładu, na tym, po której stronie osi znajdują się dane punkty.

2.5 Przykłady zastosowań

Do zaprezentowania idei i możliwości wielowymiarowej analizy korespondencji zdecydowałem się w obu przypadkach posłużyć symulowanymi danymi.

2.5.1 Przykład 1

Załóżmy, że badamy grupy wyborców dla 3 głównych partii. Posiadamy 3 zmienne, takie jak: miejsce zamieszkania (wieś, miasteczko, miasto), płeć oraz preferencje polityczne (partia A, partia B, partia C). Po przeprowadzeniu odpowiednich ankiet wycinek odpowiedzi na te pytania wygląda następująco.

	PŁEĆ	MIEJSCE	PARTIA
0	K	miasto	B
1	K	wieś	A
2	K	miasteczko	B
3	M	wieś	C
4	K	miasto	C

Rysunek 1: Przykład uzyskanych danych ankietowych

Aby wykorzystać zebrane dane, jednym ze sposobów jest je zakodować zgodnie z macierzą kodów.

	PŁEĆ_K	PŁEĆ_M	MIEJSCE_miasteczko	MIEJSCE_miasto	MIEJSCE_wieś	PARTIA_A	PARTIA_B	PARTIA_C
0	1	0	0	1	0	0	1	0
1	1	0	0	0	1	1	0	0
2	1	0	1	0	0	0	1	0
3	0	1	0	0	1	0	0	1
4	1	0	0	1	0	0	0	1

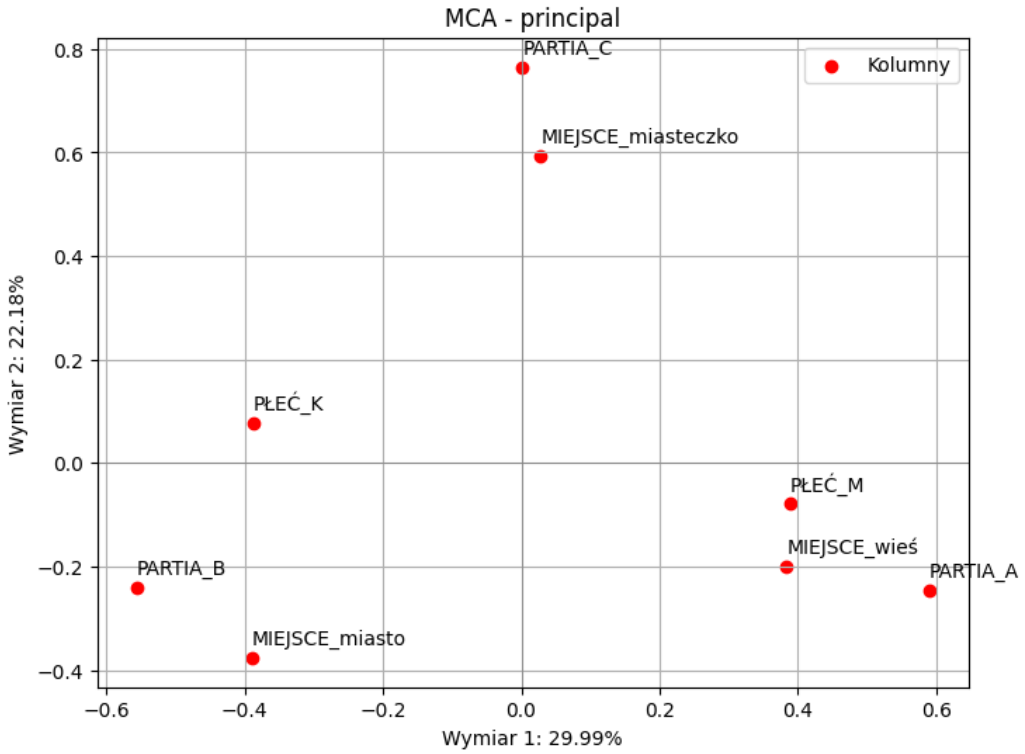
Rysunek 2: Macierz kodów dla symulowanych danych

Dane również można zapisać zgodnie z tablicą Burt'a, która dla zebranych danych prezentuje się tak jak poniżej.

	PŁEĆ_K	PŁEĆ_M	MIEJSCE_miasteczko	MIEJSCE_miasto	MIEJSCE_wieś	PARTIA_A	PARTIA_B	PARTIA_C
PŁEĆ_K	4966	0	1629	1612	1725	986	2725	1255
PŁEĆ_M	0	5034	1623	1727	1684	2799	1077	1158
MIEJSCE_miasteczko	1629	1623	3252	0	0	1121	1185	946
MIEJSCE_miasto	1612	1727	0	3339	0	787	1848	704
MIEJSCE_wieś	1725	1684	0	0	3409	1877	769	763
PARTIA_A	986	2799	1121	787	1877	3785	0	0
PARTIA_B	2725	1077	1185	1848	769	0	3802	0
PARTIA_C	1255	1158	946	704	763	0	0	2413

Rysunek 3: Tablica Burt'a dla symulowanych danych

Rekordy zapisane w jednej z dwóch wyżej wymienionych form można wykorzystać do stworzenia wykresu wielowymiarowej analizy korespondencji.



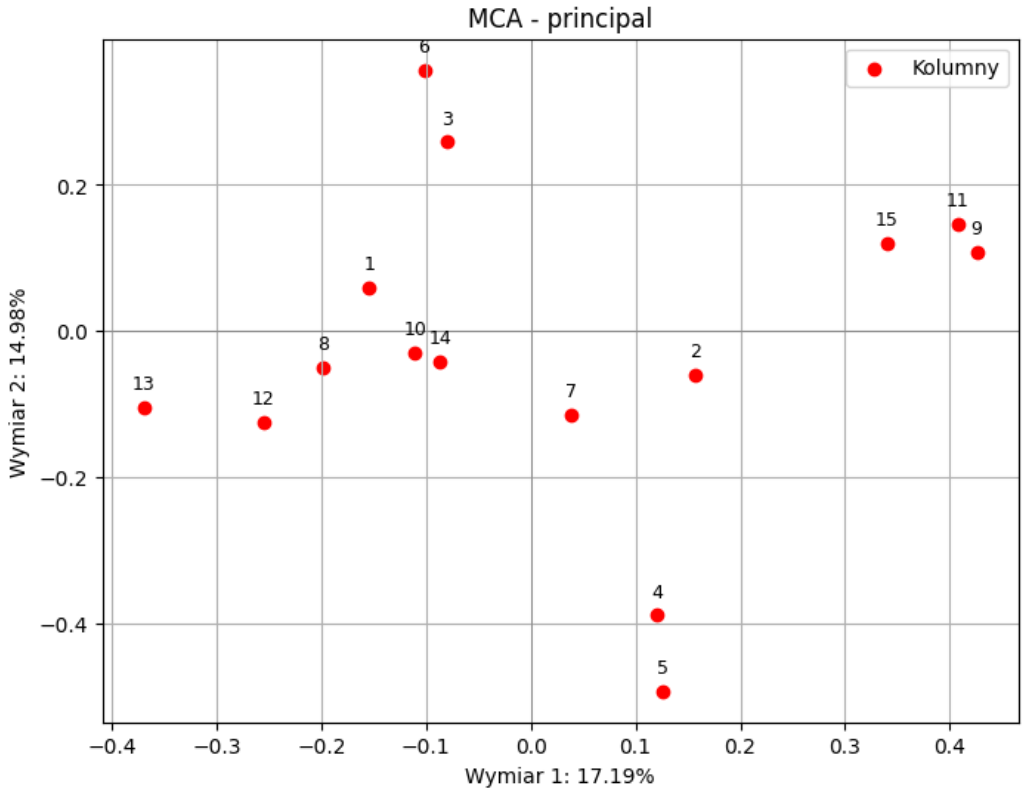
Rysunek 4: Wielowymiarowa Analiza Korespondencji

Z uzyskanego wykresu, analizując odległość między punktami opisującymi zmienne z różnych kategorii, możemy stwierdzić, czy jakieś zmienne ze sobą współwystępują. Zauważyć możemy między innymi, że partia A w dużo większym stopniu przyciąga swoimi postulatami mężczyzn oraz mieszkańców terenów wiejskich. Partia B natomiast swoim przekazem celniej trafia do kobiet oraz mieszkańców dużych miast. Skupiając się na partii C, możemy zauważyć, że ich postulaty są bardziej centrowe i tylko minimalnie bliżej znajduje się punkt reprezentujący kobiety niż mężczyzn. Widać także, że mieszkańcy miasteczek chętniej deklarują przynależność poglądową do partii C, ale również polaryzujące partie A i B są im jednakowo odległe.

2.5.2 Przykład 2

Jako przykład drugi zbadajmy respondentów pod kątem zdrowego trybu życia. W tym wypadku ankietowani deklarowali swoją płeć, aktywność fizyczną (aktywny lub siedzący tryb życia), dietę (zdrowa, niezdrowa, zwykła), czy palą, skategoryzowaną ilość snu (poniżej 6h, 6-8h, więcej niż 8h), oraz poziom stresu, jaki odczuwają (niski, umiarkowany, wysoki).

Zebrane dane ponownie posłużyły do uzyskania macierzy kodów, dzięki której uzyskano wykres wielowymiarowej analizy korespondencji.



Rysunek 5: Wielowymiarowa Analiza Korespondencji

gdzie odpowiednio liczbą przyporządkowano:

1. K, 2. M, 3. aktywny, 4. siedzący, 5. dieta - niezdrowa, 6. dieta - zdrowa, 7. dieta - zwykła, 8. nie pali, 9. pali, 10. sen 6 - 8h, 11. sen mniej niz 6h, 12. sen ponad 8h, 13. stres niski, 14. stres umiarkowany, 15. stres wysoki

Jak widzimy, wraz ze wzrostem ilości kategorii zmniejsza się przejrzystość tego rodzaju wykresów. Mimo to wciąż jesteśmy w stanie odczytać najważniejsze współwystępowania. Dla przykładu wysoki stres, sen poniżej 6h oraz deklaracja palenia wyrobów tytoniowych tworzą punkty znajdujące się bardzo blisko siebie. Oczywiście wciąż nie jesteśmy w stanie ustalić, która ze zmiennych powoduje pozostałe, ale daje to dobrą podstawę do kolejnych badań. Sprawia to także, że łatwiej przypisać kolejne cechy do badanej osoby znając jej niektóre nawyki. Zauważmy, że wiedząc, iż respondent stroni od aktywności fizycznej z większym prawdopodobieństwem jesteśmy wstanie wskazać, że również nie stosuje się do reguł zdrowego odżywiania.

3 Podsumowanie

Wielowymiarowa analiza korespondencji jest niezwykle przydatnym narzędziem w przypadku analizowania zmiennych kategorycznych. Pozwala na dostrzeżenie wzajemnych połączeń między różnymi kategoriami. Dodatkowo nie posiada ograniczenia ilości analizowanych cech, w przeciwieństwie do jej klasycznego dwuwymiarowego odpowiednika. Niemniej warto zwrócić uwagę również na wady MCA. Mimo braku ograniczenia kategorii, to przy jej większej ilości wykresy mogą stać się mało czytelne. Dodatkowo analiza ta nie pozwala stwierdzić wynikania, a jedynie wskazuje na współwystępowanie zmiennych.

4 Źródła

- https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcoran.html
- https://maxhalford.github.io/prince/mca/
- https://vgonzenbach.github.io/multivariate-cookbook/multiple-correspondence-a
html