Analiza korespondencji

Opracował:

Damian Wolański

Wprowadzenie.

Analiza korespondencji to opisowa i eksploracyjna technika analizy tablic dwudzielczych i wielodzielczych, zawierających pewne miary charakteryzujące powiązanie między kolumnami i wierszami. Otrzymywane wyniki dostarczają informacji podobnych w swej naturze do rezultatów otrzymywanych w przypadku technik Analizy czynnikowej i pozwalają na analizę struktury zmiennych jakościowych tworzących tablicę. Najczęściej spotykaną tablicą tego typu jest dwuwymiarowa tablica kontyngencji (patrz przykładowo Podstawowe statystyki lub Analiza log-liniowa).

W typowej analizie korespondencji, częstości w tablicy kontyngencji są najpierw standaryzowane w ten sposób, że oblicza się częstości względne, które zsumowane we wszystkich polach (komórkach) tablicy dają 1,0. Jednym ze sposobów ukazania celów typowej analizy jest wyrażenie częstości względnych poprzez odległości pomiędzy poszczególnymi wierszami lub kolumnami w przestrzeni o małej liczbie wymiarów. Dobrą ilustracją będzie tu prosty przykład przedstawiony poniżej. Istnieje wiele podobieństw w koncepcji i interpretacji wyników pomiędzy analizą korespondencji i analizą czynnikową , na co również zwrócimy uwagę poniżej.

Obszerny opis metody, szczegóły obliczeniowe oraz zastosowania (opublikowane w języku angielskim) można znaleźć w klasycznej pracy Greenacre'a (1984). Metody analizy korespondencji rozwinięte zostały głównie we Francji we wczesnych latach sześćdziesiątych, a potem siedemdziesiątych przez Jean-Paula Benzérci (np. patrz Benzérci, 1973 lub Lebart, Morineau, Tabard, 1977) i dopiero ostatnio zdobyły popularność w krajach anglojęzycznych (patrz np. Carrol, Green, Schaffer, 1986; Hoffman i Franke, 1986). (Warto zauważyć, że podobne techniki rozwijano niezależnie w kilku krajach, gdzie znane są pod takimi nazwami, jak skalowanie optymalne, wzajemne uśrednianie, optymalne punktowanie, metoda kwantyfikacji lub też analiza jednorodności). W kolejnych akapitach przedstawione zostanie ogólne wprowadzenie w problematykę analizy korespondencji.

Informacje wstępne. Załóżmy, że zebraliśmy dane dotyczące intensywności palenia papierosów wśród pracowników pewnej firmy. Przedstawiony tu zbiór danych został zaczerpnięty z pracy Greenacre'a (1984, str. 55).

	Intensywność palenia				
Grupa pracownicza	**(1) Nie pali**	**(2) Mało**	**(3) Średnio**	**(4) Dużo**	**Suma wierszy**
(1) Starsi kierownicy	4	2	3	2	11
(2) Młodsi kierownicy	4	3	7	4	18
(3) Starszy personel	25	10	12	4	51
(4) Młodszy personel	18	24	33	13	88
(5) Sekretarki	10	6	7	2	25
Sumy kolumn	61	45	62	25	193

Każde cztery wartości w poszczególnych wierszach można potraktować jako współrzędne w przestrzeni czterowymiarowej i można policzyć odległości (euklidesowe) pomiędzy pięcioma punktami (wierszami) w tej przestrzeni. Odległości te zawierają sumaryczną informację o podobieństwie wierszy powyższej tablicy. Teraz załóżmy, że chcemy znaleźć przestrzeń o mniejszej (od 4) liczbie wymiarów, w której da się umieścić punkty odpowiadające wierszom w ten sposób, że zachowana zostanie pełna lub prawie pełna informacja o zróżnicowaniu wierszy. Wówczas można tę informację o podobieństwach między wierszami (grupami pracowników) przedstawić na prostym jedno-, dwu- lub trójwymiarowym rysunku. Procedura taka może wydawać się niezbyt przydatną dla małych tablic, takich jak ta przedstawiona w powyższym przykładzie, jednakże łatwo można sobie wyobrazić, jak zastosowanie analizy korespondencji ułatwia prezentację i interpretację bardzo dużych tablic (np. zróżnicowanie preferencji dotyczących 10 dóbr konsumpcyjnych wśród 100 grup respondentów przedstawione poprzez pozycje 10 dóbr w przestrzeni dwuwymiarowej).

Masa.

Kontynuując prosty przykład zaprezentowany w tabeli, program po pierwsze obliczy częstości względne, które w całej tablicy kontyngencji sumują się do 1,0 (każdą liczebność empiryczną dzieli się przez liczebność ogólną, czyli tutaj *193*). Można powiedzieć, że częstości względne pokazują, jak jednostka *masy* jest rozłożona na poszczególne komórki. W terminologii używanej w analizie korespondencji sumy częstości względnych dla poszczególnych wierszy i kolumn nazywane są odpowiednio masą wiersza i masą kolumny.

Bezwładność.

Termin bezwładność jest używany w analizie korespondencji analogicznie do występującego w matematyce stosowanej pojęcia momentu bezwładności , który definiowany jest jako całka masy pomnożonej przez kwadrat odległości od środka ciężkości (np. Greenacre, 1984, str. 35). Bezwładność definiowana jest jako iloraz statystyki Chi-kwadrat Pearsona obliczonej z tablicy dwudzielczej (można ja obliczyć również w modułach Podstawowe statystyki oraz Analiza log-liniowa) przez liczebność ogólną (w przedstawionym przykładzie liczebność ogólna wynosi *193*).

Bezwładność oraz profile wierszy i kolumn.

Jeżeli wiersze i kolumny w tablicy są całkowicie niezależne, to wartości w komórkach tej tablicy (rozkład masy) można wyznaczyć na podstawie wyłącznie sum wierszy i kolumn, które to sumy są w terminologii analizy korespondencji nazywane *profilami*. Zgodnie ze znanym wzorem na obliczanie statystyki Chi-kwadrat z tablicy dwudzielczej, częstości oczekiwane oblicza się jako iloczyny odpowiednich liczebności brzegowych podzielone przez liczebność ogólną przy założeniu niezależności cech reprezentowanych przez wiersze i kolumny. Każde odstępstwo od wartości oczekiwanych (oczekiwanych przy prawdziwości hipotezy o niezależności) wpływa na ogólną wartość statystyki *Chi-kwadrat*. Tak więc inny sposób spojrzenia na analizę korespondencji to potraktowanie jej jako metody dekompozycji ogólnej statystyki *Chi-kwadrat* (lub Bezwładność = *Chi-kwadrat/N*) poprzez zdefiniowanie układu o małej liczbie wymiarów, w którym zaprezentuje się odchylenia od wartości oczekiwanych. Mamy tu podobieństwo do Analizy czynnikowej , gdzie całkowita wariancja jest dekomponowana aż do uzyskania reprezentacji zmiennych w przestrzeni o małej liczbie wymiarów w taki sposób, aby jak najtrafniej móc z tej reprezentacji odtworzyć oryginalną macierz wariancji/kowariancji zmiennych.

Analiza wierszy i kolumn.

Przedstawiony przykład empiryczny rozpoczęliśmy od rozważania punktów reprezentujących wiersze w przytoczonej tablicy. Można oczywiście zainteresować się bardziej sumami kolumn i starać się umieścić punkty reprezentujące kolumny w przestrzeni o małej liczbie wymiarów, która pozwoli na zadowalające przedstawienie podobieństw (lub odległości) pomiędzy częstościami względnymi zawartymi w kolumnach. W praktyce, punkty reprezentujące wiersze i kolumny umieszcza się zwykle na wspólnym wykresie, który jest pewnym sumarycznym przedstawieniem informacji zawartych w tablicy dwudzielczej.

Przeglądanie wyników.

Przyjrzyjmy się niektórym wynikom zawartym w tabeli poniżej. Mamy tu tak zwane *wartości osobliwe, wartości własne, procenty wyjaśnionej bezwładności , procenty skumulowane* i wkład w ogólną wartość statystyki *Chi-kwadrat*.

Wartości własne i Bezwładność dla wszystkich wymiarów
Tablica wejściowa (Wiersze x Kolumny): 5 x 4
Całkowita bezwładność = .08519 Chi2 = 16.442

Nr wymiaru	Wartości osobliwe	Wartości własne	Procent bezwładności	Procent skumulowany	Chi kwadrat
1	.273421	.074759	87.75587	87.7559	14.42851
2	.100086	.010017	11.75865	99.5145	1.93332
3	.020337	.000414	.48547	100.0000	.07982

Zauważmy, że wymiary są "wyznaczane" w taki sposób, aby maksymalizować odległości między punktami reprezentującymi wiersze lub kolumny, a kolejne wymiary (które są

niezależne lub ortogonalne względem siebie) "wyjaśniają" coraz mniejsze części ogólnej *Chi-kwadrat* (czyli również bezwładności . Jak widać wyodrębnianie kolejnych wymiarów jest podobne do wyznaczania *głównych składowych* w *Analizie czynnikowej* .

Widzimy, że przy jednym wymiarze, może być wyjaśnione 87.76% bezwładności , co oznacza, że częstości względne, które można odtworzyć z informacji jednowymiarowych mogą odtworzyć 87.76% ogólnej wartości statystyki *Chi-kwadrat* (a więc także *bezwładności*) dla tej tablicy dwudzielczej; dwa wymiary pozwalają wyjaśnić 99.51%.

Maksymalna liczba wymiarów. Ponieważ sumy częstości w wierszach i kolumnach nie mogą ulegać zmianie, zatem w pewnym sensie w każdym wierszu mamy (*liczba kolumn - 1*) niezależnych wartości, zaś w każdej kolumnie jest (*liczba wierszy - 1*) niezależnych wartości (znając je można wyliczyć resztę tablicy wykorzystując liczebności brzegowe). W związku z tym największa liczba wartości własnych, które mogą zostać wyznaczone z tablicy dwudzielczej jest równa liczbie wierszy pomniejszonej o 1, jeżeli w tablicy liczba wierszy jest nie większa niż liczba kolumn, lub liczbie kolumn minus 1, gdy kolumn jest mniej niż wierszy. Gdy zdecydujemy się na wyodrębnienie (interpretację) maksymalnej liczby wymiarów, które można uwzględnić, wtedy możemy odtworzyć dokładnie pełną informację zawartą w początkowej tablicy kontyngencji.

Współrzędne wierszy i kolumn. Spójrzmy teraz na współrzędne uzyskane w rozwiązaniu dwuwymiarowym.

Nazwa wiersza	Wymiar 1	Wymiar 2
(1) Starsi kierownicy	-.065768	.193737
(2) Młodsi kierownicy	.258958	.243305
(3) Starszy personel	-.380595	.010660
(4) Młodszy personel	.232952	-.057744
(5) Sekretarki	-.201089	-.078911

Oczywiście teraz można wykreślić punkty o tych współrzędnych na dwuwymiarowym wykresie rozrzutu. Pamiętajmy, że celem analizy korespondencji jest odtworzenie odległości pomiędzy punktami reprezentującymi wiersze lub kolumny tablicy dwudzielczej, w przestrzeni o mniejszej liczbie wymiarów. Zauważmy, że podobnie jak w *Analizie czynnikowej* kierunki kolejnych osi (wymiarów) ustalonych tak, że kolejne wymiary wyjaśniają coraz mniejszą część ogólnej wartości *Chi-kwadrat* (lub bezwładności), są dobierane arbitralnie. W każdej kolumnie tablicy przedstawionej powyżej można zmienić znaki na przeciwne, co jest równoznaczne z rotacją danej osi o 180°.

Odległości punktów na wykresie dwuwymiarowym dostarczają informacji o podobieństwie częstości względnych, jakie dane wiersze mają w odpowiednich kolumnach. Na rysunku ilustrującym nasz przykład, grupy *Starszy personel* i *Sekretarki* są relatywnie blisko siebie, jeżeli brać pod uwagę pierwszą, najważniejszą oś, po lewej stronie od jej punktu początkowego (zera). W tablicy częstości względnych standaryzowanych osobno dla każdego wiersza możemy zauważyć, że struktury tych grup pracowników według intensywności palenia tytoniu są rzeczywiście bardzo podobne.

Rozkłady procentowe dla wierszy

	Intensywność palenia				
Grupa pracownicza	**(1) Nie pali**	**(2) Mało**	**(3) Średnio**	**(4) Dużo**	**Sumy wierszy**
(1) Starsi kierownicy	36.36	18.18	27.27	18.18	100.00
(2) Młodsi kierownicy	22.22	16.67	38.89	22.22	100.00
(3) Starszy personel	49.02	19.61	23.53	7.84	100.00
(4) Młodszy personel	20.45	27.27	37.50	14.77	100.00
(5) Sekretarki	40.00	24.00	28.00	8.00	100.00

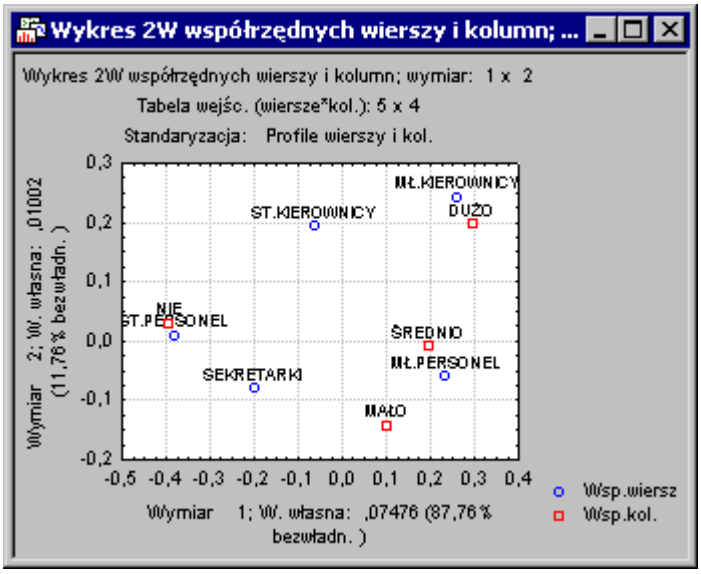
Ostatecznym celem analizy korespondencji jest oczywiście znalezienie teoretycznej interpretacji (znaczenia) wyodrębnionych wymiarów. Jedną z metod, która może być pomocna w tej interpretacji jest sporządzenie wykresu punktów odpowiadających kolumnom. Poniżej podano współrzędne kolumn dla pierwszego i drugiego wymiaru.

Intensywność palenia	**Wymiar 1**	**Wymiar 2**
Nie pali	-.393308	.030492
Mało	.099456	-.141064
Średnio	.196321	-.007359
Dużo	.293776	.197766

Wydaje się, że pierwszy wymiar najbardziej różnicuje różne kategorie pracowników ze względu na intensywność palenia, a szczególne kategorię Nie pali od pozostałych kategorii. Można więc interpretować duże podobieństwo współrzędnych pierwszego wymiaru dla grup *Starsi kierownicy* i *Sekretarki* jako wynikające głównie ze stosunkowo dużej liczby niepalących w tych grupach.

Porównywalność współrzędnych wierszy i kolumn.

Często przedstawia się wiersze i kolumny na jednym wykresie. Trzeba jednak pamiętać, że na takim wspólnym wykresie można interpretować tylko odległości między punktami reprezentującymi wiersze, albo odległości między punktami reprezentującymi kolumny, ale nie odległości między kolumnami a wierszami.



Pozostając przy rozważanym przykładzie, nie możemy przecież twierdzić, że kategoria *Nie* pali jest podobna do kategorii *Starszy personel* (mimo, że te punkty są bardzo blisko na wspólnym wykresie kolumn i wierszy). Jednakże można, jak to wskazano uprzednio, formułować pewne ogólne spostrzeżenia na temat charakteru wymiarów (osi), oparte na stwierdzeniu, po której stronie osi znajdują się dane punkty. Dla przykładu, ponieważ punkt reprezentujący kategorię *Nie* pali jest jedynym punktem znajdującym się po lewej stronie pierwszej osi, a grupa *Starsi kierownicy* także znajduje się po tej samej stronie względem początku pierwszej osi, to można wnioskować, że pierwsza oś pozwala odróżnić kategorię *Nie* pali od różnych kategorii palaczy oraz, że *Starsi kierownicy* różnią się na przykład od *Młodszego personelu* tym, że wśród *Starszych kierowników* jest relatywnie więcej niepalących.

Skalowanie współrzędnych (metody standaryzacji).

Kolejna ważna decyzja, którą musi podjąć badacz dotyczy skalowania współrzędnych. Należy zdecydować, czy analizuje się relatywne częstości procentowe (charakteryzujące tak zwane rozkłady warunkowe) dla wierszy, dla kolumn czy też obydwa ujęcia. W omawianym przykładzie przytoczono rozkłady procentowe dla wierszy, aby pokazać jak rozkłady te są podobne dla punktów (reprezentujących wiersze) leżących blisko siebie na wykresie. Mówiąc nieco inaczej, współrzędne punktów wyznacza się w oparciu o *macierz profili wierszy*, w której suma każdego wiersza jest równa jedności (każdy element r_{ij} macierzy profili wierszy może być interpretowany jako prawdopodobieństwo warunkowe tego, że obserwacja należy do kolumny *j*, przy założeniu, że należy do wiersza i). Współrzędne są tak wyznaczane, aby

maksymalizować różnice między punktami z punktu widzenia profili wierszy (rozkładów procentowych wierszy). Współrzędne wierszy są wyliczane z macierzy profili wierszy a współrzędne kolumn są wyliczane z macierzy profili kolumn.

Mamy również czwartą możliwość, standaryzację *kanoniczną*, która polega na standaryzacji kolumn i wierszy macierzy częstości względnych (patrz Gifi, 1981). Omawiany typ standaryzacji, rzadko stosowany, sprowadza się do przeskalowania współrzędnych uzyskanych po standaryzacji wierszowej oraz standaryzacji kolumnowej. Zauważmy, że łatwo można zastosować inne typy standaryzacji, jeżeli mamy nie przekształcone wartości własne oraz macierze wektorów własnych.

Metryka układu współrzędnych.

W wielu miejscach tego wprowadzenia, pojęcie odległości było używane w *odniesieniu* do różnic pomiędzy strukturami częstości względnych wierszy oraz pomiędzy strukturami częstości względnych kolumn, które w rezultacie zastosowania analizy korespondencji mają być odtworzone w przestrzeni o liczbie wymiarów mniejszej niż oryginalna przestrzeń danych. W rzeczywistości odległości te nie są prostymi odległościami euklidesowymi obliczonymi z częstości względnych wierszy i kolumn, ale są to odległości ważone. Stosowane wagi są tego typu, że taka metryka w przestrzeni o mniejszej liczbie wymiarów jest metryką typu *Chi-kwadrat*, przy założeniu, że (1) przy porównywaniu punktów reprezentujących wiersze wybraliśmy standaryzację wierszową (profile wierszy) lub wierszowo-kolumnową, zaś przy (2) porównywaniu punktów reprezentujących kolumny standaryzacja jest typu kolumnowego (profile kolumn) lub wierszowo-kolumnowego (profile wierszy i kolumn).

W takim przypadku (*nie*odnosi się to wszakże do standaryzacji kanonicznej), kwadrat odległości euklidesowej pomiędzy, przykładowo, dwoma punktami reprezentującymi wiersze *i* oraz *i'* w odpowiednim układzie współrzędnych o danej liczbie wymiarów jest w przybliżeniu równy ważonej (tj. Chi-kwadrat) odległości pomiędzy częstościami względnymi (patrz Hoffman i Franke, 1986, wzór 21):

$$d_{ii'}^2 = \Sigma_j \left(1/c_j \left(p_{ij}/r_i - p^2_{i'j}/r_{i'}\right)\right)$$

W tym wzorze $d_{ii'}$ to kwadrat odległości między dwoma punktami, c_j jest sumą *j*-tej kolumny w tablicy częstości standaryzowanych (w której suma wszystkich wartości lub masy jest równa *1,0*), p_{ij} to częstości standaryzowane (*i* jest numerem wiersza, a *j*) numerem kolumny), r_i jest sumą częstości względnych *i*-tego wiersza, a sumowanie Σprzebiega po wszystkich kolumnach tablicy. Powtórzmy, że interpretowalne są odległości tylko między wierszami lub tylko między kolumnami, a nie pomiędzy wierszami a kolumnami.

Ocena jakości rozwiązania.

Podawanych jest kilka dodatkowych statystyk, które są pomocne przy ocenie jakości rozwiązania o wybranej liczbie wymiarów. Ogólnie chodzi o to, czy wszystkie (lub przynajmniej większość) punkty są właściwie reprezentowane, to znaczy czy ich odległości do innych punktów są aproksymowane z zadowalającą dokładnością. Poniżej zaprezentowano wszystkie statystyki dotyczące omawianego przykładu, wyliczone dla współrzędnych wierszy

na podstawie rozwiązania w przestrzeni o tylko jednym wymiarze (użyto więc tylko jednej osi do odtworzenia częstości względnych w kolumnach).

Współrzędne wierszy i udziały w bezwładności

Grupa pracownicza	Współrz. wymiar 1	Masa	Jakość	Bezwładność względna	Bezwładność wymiar 1	Cos² wymiar 1
(1) Starsi kierownicy	-.065768	.056995	.092232	.031376	.003298	.092232
(2) Młodsi kierownicy	.258958	.093264	.526400	.139467	.083659	.526400
(3) Starszy personel	-.380595	.264249	.999033	.449750	.512006	.999033
(4) Młodszy personel	.232952	.455959	.941934	.308354	.330974	.941934
(5) Sekretarki	-.201089	.129534	.865346	.071053	.070064	.865346

Współrzędne.

Pierwsza kolumna liczb w powyższej tabeli zawiera współrzędne omawiane w poprzednich akapitach. Przypomnijmy, że interpretacja tych współrzędnych zależy od typu wybranej standaryzacji. Liczba wymiarów wybierana jest przez użytkownika (w omawianym przykładzie wybraliśmy tylko jeden wymiar) i współrzędne podawane są dla każdego wymiaru (dla każdego wymiaru jest osobna kolumna).

Masa. Kolumna *Masa* zawiera sumy wierszy (gdyż mamy do czynienia ze współrzędnymi wierszy) obliczone z tablicy częstości względnych (czyli z tablicy, której każdy element oznacza odpowiednią *Masę*, jak to wyjaśniono wcześniej). Mówiąc inaczej, współrzędne wyliczane są z macierzy prawdopodobieństw warunkowych pokazywanych w kolumnie *Masa*.

Jakość . Kolumna *Jakość* zawiera informacje dotyczące jakości reprezentacji wiersza przez punkt w wybranym układzie współrzędnych. W tablicy przedstawionej powyżej, wybrano tylko jeden wymiar, wobec tego liczby w kolumnie Jakość dotyczą jakości reprezentacji w przestrzeni jednowymiarowej. Powtórzmy, że z rachunkowego punktu widzenia celem analizy korespondencji jest odtworzenie odległości między punktami w przestrzeni o mniejszej liczbie wymiarów. Gdy zdecydujemy się na utworzenie układu z maksymalną liczbą wymiarów (która jest równa mniejszej z dwóch liczb: liczba kolumn pomniejszona o jeden oraz liczba wierszy pomniejszona o jeden), wtedy można dokładnie odtworzyć dokładnie wszystkie oryginalne odległości. Jakość punktu jest definiowana jako iloraz kwadratu odległości danego punktu od środka wybranego układu współrzędnych przez kwadrat takiej odległości w układzie współrzędnych o maksymalnej liczbie wymiarów (pamiętając, że stosujemy tu metrykę typu *Chi-kwadrat*). Interpretacja jakości punktu w analizie korespondencji jest analogiczna do pojęcia zasobu zmienności wspólnej w *analizie czynnikowej* .

Zauważmy, że wartość miary *jakości* wyliczamy nie zależy od wyboru metody standaryzacji i zawsze wynika z domyślnej standaryzacji (stosuje się tu metrykę odległości typu *Chi-kwadrat*, a miara jakości może być interpretowana jako część ogólnej wartości statystyki Chi-kwadrat odpowiadająca danemu wierszowi, przy danej liczbie wymiarów. Niska jakość

oznacza, że rozpatrywana właśnie liczba wymiarów nie pozwala na dobrą reprezentację danego wiersza (lub kolumny). W powyższej tabeli jakość pierwszego wiersza (*Starsi kierownicy*) jest mniejsza od *0,1*, co oznacza, że punkt w przestrzeni jednowymiarowej niebyt dobrze reprezentuje oryginalną informację znajdującą się w pierwszym wierszu.

Bezwładność względna.

Jakość punktu (patrz wyżej) wyraża udział danego punktu w ogólnej bezwładności (*Chi-kwadrat*), przy danym wymiarze przestrzeni. Jednakże nie wynika stąd, czy i w jakim zakresie odpowiedni punkt wpływa na ogólną bezwładność (*Chi-kwadrat*) w wyjściowym układzie współrzędnych. Bezwładność względna wyraża udział danego punktu w bezwładności ogólnej mierzony niezależnie od liczby wymiarów wybranych przez badacza. Zauważmy, że konkretne rozwiązanie może odwzorowywać dany punkt bardzo dobrze (wysoka *jakość*), ale ten sam punkt może nie mieć dużego wpływu na bezwładność ogólną (przykładem może być wiersz, którego częstości względne są bardzo zbliżone do rozkładu brzegowego).

Bezwładność względna dla każdego wymiaru. Kolumna ta zawiera względne udziały poszczególnych wierszy w bezwładności "generowanej przez" dany wymiar, tak więc wartość ta będzie podawana osobno dla każdego wymiaru.

Cos^2 (jakość lub kwadrat korelacji z każdym wymiarem). Ta kolumna zawiera miary jakości dla każdego punktu w odniesieniu do każdego wymiaru. Suma wartości tych kolumn liczona po wszystkich wymiarach jest równa całkowitej *jakości* omówionej powyżej (ponieważ w rozważanym przykładzie wybrano tylko jeden wymiar, dlatego wartości w tej kolumnie są takie same jak wartości zamieszczone w kolumnie *Jakość*). Omawiana wartość może być również interpretowana jako korelacja danego punktu z odpowiednim wymiarem. *Kwadrat cosinusa* wiąże się z faktem, że wartość ta jest również równa kwadratowi cosinusa kąta pomiędzy wektorem określonym przez dany punkt a odpowiednią osią układu (szczegóły geometrycznych aspektów analizy korespondencji znajdują się w pracy Greenacrea, 1984).

Uwaga na temat "istotności statystycznej". W tym miejscu należy stwierdzić, że analiza korespondencji jest techniką eksploracyjną. Tak naprawdę w rozwoju tej metody położono nacisk bardziej na poszukiwanie modeli, które dobrze opisują dane empiryczne, niż na odrzucenie hipotez dotyczących braku dopasowania (patrz "druga zasada" Benzecriego głosząca, że "To model ma pasować do danych, a nie na odwrót", Greenacre 1984, str.10). W związku z tym nie ma statystycznych testów istotności, które zwyczajowo stosowałoby się do wyników analizy korespondencji. Pierwotnym celem tej techniki jest stworzenie uproszczonego (w przestrzeni o małej liczbie wymiarów) odwzorowania informacji zawartej w dużej tablicy kontyngencji (lub analogicznych tablicach zawierających miary związku między wariantami cech).

Punkty dodatkowe

W części Wprowadzenie zamieszczono wskazówki, jak należy interpretować współrzędne oraz związane z nimi statystyki uzyskane w analizie korespondencji. W trakcie interpretacji wyników bardzo pomocne może być dodanie wiersza lub kolumny, które nie były wykorzystywane w dotychczasowej analizie. Ilustracją niech będzie przykład empiryczny rozważany już w części Wprowadzenie (przykład został wzięty z pracy Greenacre'a, 1984).

Grupa pracownicza	Wymiar 1	Wymiar 2
(1) Starsi kierownicy	-.065768	.193737
(2) Młodsi kierownicy	.258958	.243305
(3) Starszy personel	-.380595	.010660
(4) Młodszy personel	.232952	-.057744
(5) Sekretarki	-.201089	-.078911
Średnia krajowa	-.258368	-.117648

Powyższa tablica pokazuje wartości współrzędnych (w przestrzeni dwuwymiarowej) wyliczone z tablicy kontyngencji charakteryzującej grupy pracownicze w zależności od intensywności palenia papierosów. Wiersz o nazwie *Średnia krajowa* zawiera współrzędne dodatkowego punktu, określającego intensywność palenia dla całej populacji Stanów Zjednoczonych. (Za pracą Greenacre'a przyjmujemy następujące dane fikcyjne: niepalący: 42%, palący mało: 29%, palący średnio: 20%, palący dużo: 9%). Na dwuwymiarowym wykresie rozrzutu zauważamy, że punkt reprezentujący *Średnią krajową* znajduje się blisko punktu Sekretarki oraz że znajduje się on po tej samej stronie osi poziomej (wymiar pierwszy) co punkt reprezentujący kolumnę *Niepalący*. Te spostrzeżenia są zgodne z wyjściowymi danymi zawartymi w tablicy dwudzielczej zamieszczonej w części Wprowadzenie . Analizując rozkłady procentowe w wierszach zauważamy, że relatywnie więcej niepalących jest w wierszach *Sekretarki* i *Średnia krajowa*. Mówiąc inaczej, w próbie reprezentowanej przez wspomnianą tablicę częstości jest relatywnie więcej palaczy niż w populacji ogólnej.

O ile taka informacja może być łatwo wydedukowana z tablicy częstości o małych rozmiarach (takiej jak w omawianym przykładzie), o tyle przy dużych tablicach niełatwo bezpośrednio zauważyć prawidłowości.

Jakość odwzorowania punktów dodatkowych. Dalsze interesujące wyniki analizy związane z punktami dodatkowymi dotyczą jakości ich odwzorowania w układzie współrzędnych o wybranej liczbie wymiarów (w części Wprowadzenie omówiono bliżej pojęcie *jakości odwzorowania*). Przypomnijmy, że celem analizy korespondencji jest odtworzenie odległości między wierszami lub kolumnami w układzie współrzędnych o niewielkiej liczbie osi. Mając takie rozwiązanie, można zapytać, czy konkretny punkt dodatkowy jest dobrze odwzorowany w przestrzeni wynikowej, to znaczy czy jego odległości do innych punktów są dobrze odwzorowane w przestrzeni o wybranej liczbie wymiarów. Poniżej przedstawiamy wartości statystyk dotyczące punktów wyjściowych przykładu oraz wiersza dodatkowego *Średnia krajowa* dla rozwiązania w przestrzeni dwuwymiarowej.

Grupa pracownicza	Jakość	Cos² wymiar 1	Cos² wymiar 2
(1) Starsi kierownicy	.892568	.092232	.800336
(2) Młodsi kierownicy	.991082	.526400	.464682
(3) Starszy personel	.999817	.999033	.000784
(4) Młodszy personel	.999810	.941934	.057876
(5) Sekretarki	.998603	.865346	.133257
Średnia krajowa	.761324	.630578	.130746

Statystyki podane w powyższej tablicy są omówione w części Wprowadzenie . *Jakość* punktu reprezentującego wiersz lub kolumnę jest definiowana jako iloraz kwadratu odległości danego punktu od początku wynikowego układu współrzędnych do kwadratu odległości od początku

układu zawierającego maksymalną liczbę wymiarów (pamiętajmy, że stosuje się tu metrykę typu *Chi-kwadrat* - patrz Wprowadzenie). W tym sensie, ogólna jakość określa jaka część sumy kwadratów odległości od środka ciężkości jest wyjaśniona w wynikowym układzie współrzędnych. Punkt reprezentujący dodatkowy wiersz *Średnia krajowa* charakteryzuje się jakością *0,76*, co oznacza, że jest dość dobrze odwzorowany w przestrzeni dwuwymiarowej. Statystyka *Cosinus*^{2} określa *jakość* odwzorowania danego punktu dotyczącą konkretnego wymiaru (suma wartości statystyki Cosinus^2 dzielona przez odpowiednia liczbę wymiarów jest równa całkowitej *Jakości* - patrz Wprowadzenie).

Wielowymiarowa analiza korespondencji (MCA)

Wielowymiarowa analiza korespondencji (Multiple Correspondence Analysis - MCA) może być rozpatrywana jako rozszerzenie prostej analizy korespondencji na zagadnienia o liczbie zmiennych większej od dwóch. Zagadnienia wstępne prostej analizy korespondencji przedstawiono we Wprowadzeniu . Wielowymiarowa analiza korespondencji to prosta analiza korespondencji prowadzona na macierzy kodów (układu), gdzie poszczególne wiersze odpowiadają kolejnym obserwacjom, a kolumny - wariantom zmiennych. W rzeczywistości, zazwyczaj analizuje się iloczyn wewnętrzny takiej macierzy, nazywany w wielowymiarowej analizie korespondencji *Tablicą Burta* . Jednakże dla wytłumaczenia interpretacji wyników wielowymiarowej analizy korespondencji, łatwiej jest odwołać się do prostej analizy korespondencji przeprowadzanej na macierzy kodów (układu).

Macierz kodów (układu). Rozważmy ponownie prostą tablicę dwudzielczą przedstawioną już w części Wprowadzenie :

	Intensywność palenia				
Grupa pracownicza	**(1) Nie pali**	**(2) Mało**	**(3) Średnio**	**(4) Dużo**	**Sumy wierszy**
(1) Starsi kierownicy	4	2	3	2	11
(2) Młodsi kierownicy	4	3	7	4	18
(3) Starszy personel	25	10	12	4	51
(4) Młodszy personel	18	24	33	13	88
(5) Sekretarki	10	6	7	2	25
Sumy kolumn	61	45	62	25	193

Załóżmy, że dane do tej tablicy zostały wprowadzone w sposób podany poniżej, jako macierz *kodów* (*układu*):

	Grupa pracownicza					Intensywność palenia			
Nr obserwacji	**Starszy kierownik**	**Młodszy kierownik**	**Starszy personel**	**Młodszy personel**	**Sekretarka**	**Nie pali**	**Mało**	**Średnio**	**Dużo**
1	1	0	0	0	0	1	0	0	0
2	1	0	0	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	0
4	1	0	0	0	0	1	0	0	0
5	1	0	0	0	0	0	1	0	0
...	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.
191	0	0	0	0	1	0	0	1	0
192	0	0	0	0	1	0	0	0	1
193	0	0	0	0	1	0	0	0	1

Każda z 193 obserwacji w tablicy jest reprezentowana przez jeden wiersz w pliku danych. Dla każdej obserwacji (przypadku) zaznaczamy *1* w kategorii (wariancie cechy), do której dany przypadek należy, natomiast w pozostałych kategoriach stawiamy *0*. Przykładowo, obserwacja numer *1*, to Starszy kierownik, *Niepalący*. Jak widać w tabeli dwudzielczej, mamy łącznie cztery takie przypadki, wobec tego w macierzy kodów będą cztery wiersze o postaci takiej, jak wiersz pierwszy. Macierz kodów (układu) odpowiadająca omawianemu przykładowi zawiera *193* wiersze.

Analiza macierzy kodów (układu). Jeżeli macierz kodów przedstawioną powyżej poddamy analizie, tak jakby to była zwykła tablica dwudzielcza, wtedy wyniki analizy korespondencji dostarczą nam współrzędnych kolumn, które pozwolą z kolei na analizę relacji między poszczególnymi kategoriami, wynikających z odległości pomiędzy poszczególnymi wierszami (przypadkami, obserwacjami). Wykres dwuwymiarowy jaki wówczas otrzymamy będzie bardzo podobny do wykresu obrazującego łączne położenie punktów dla kolumn i wierszy uzyskanego z prostej analizy korespondencji przeprowadzonej na tablicy dwudzielczej (zauważmy, że metryka jest tu odmienna, ale wzajemne położenie punktów bardzo podobne).

Przypadek więcej niż dwóch zmiennych. Przedstawiony powyżej sposób analizy cech jakościowych może być łatwo uogólniony na zagadnienia, w których obserwacje opisane są przy pomocy więcej niż dwu cech jakościowych. Na przykład, do rozpatrywanych danych możemy dodać dwie zmienne: *Mężczyzna* i *Kobieta*. W macierzy kodów przybędą wówczas dwie kolumny dla zakodowania Płci *0* i *1* oraz (powiedzmy) trzy kolumny do oznaczenia przynależności danej osoby do jednej z trzech grup wiekowych. Zatem w końcowym efekcie będziemy mogli przedstawić relacje (podobieństwo) pomiędzy *Płcią*, *Wiekiem*, *Intensywnością palenia* i *Zawodem (Grupą pracowniczą)*.

Kodowanie rozmyte. Nie jest konieczne, aby każda obserwacja była przyporządkowana tylko do jednej kategorii w obrębie każdej z cech jakościowych. Zamiast kodowania w konwencji *0* lub *1*, możemy wpisywać prawdopodobieństwo przynależności do danej kategorii rozważanej cechy lub wartość jakiejś innej miary określającej rozmytą regułę przynależności do danej grupy. W pracy Greenacre'a (1984) omówiono kilka sposobów kodowań tego typu. Załóżmy, że w omawianym przykładzie, w odniesieniu do kilku osób nie posiadamy informacji dotyczących palenia papierosów. Zamiast usuwania tych przypadków z dalszej analizy (lub tworzenia nowej kategorii *Brak danych*), można każdej kategorii określającej intensywność palenia przyporządkować ułamek (suma tych liczb dla każdej zmiennej i każdej obserwacji ma dawać 1,0) odpowiadający prawdopodobieństwu tego, że obserwacja należy do danej kategorii (na przykład można tu wpisać wartości wynikające z ogólnokrajowych szacunków rozkładu cechy w populacji).

Interpretacja współrzędnych i pozostałych wyników. Powtórzmy, że wyniki wielowymiarowej analizy korespondencji są takie same jak współrzędne kolumn uzyskane z przeprowadzenia prostej analizy korespondencji na podstawie macierzy kodów (macierzy układu). W związku z tym, wartości współrzędnych, miary jakości , cosinusów^2 oraz innych statystyk podanych jako wyniki wielowymiarowej analizy korespondencji mogą być interpretowane w ten sam sposób jak to opisano w prostej analizie korespondencji (patrz Wprowadzenie). Należy jednak pamiętać, że te statystyki określają udział danego elementu w całkowitej bezwładności określonej na podstawie całej macierzy kodowej.

Dodatkowe punkty reprezentujące kolumny oraz "regresja wielokrotna" dla zmiennych jakościowych. Inne zastosowanie analizy macierzy układu za pomocą technik analizy korespondencji umożliwia przeprowadzenie analizy analogicznej do *regresji wielokrotnej* dla zmiennych jakościowych, poprzez wprowadzanie dodatkowych kolumn do macierzy układu. Przykładowo załóżmy, że do macierzy podanej uprzednio dodaliśmy dwie kolumny określające, czy dana osoba chorowała w ciągu minionego roku (jedna z nich może się nazywać *Chorował*, a druga *Nie chorował*, wobec czego stan zdrowia każdej osoby zakodowany jest poprzez odpowiednie umieszczenie *0* i *1* w tych kolumnach). Jeżeli wprowadziliśmy takie kolumny do macierzy układu, jako kolumny dodatkowe, wówczas (1) sumaryczne statystyki jakości odwzorowania (patrz Wprowadzenie) dotyczące tych kolumn dostarczają informacji o tym, w jakim stopniu fakt zachorowania może być "wyjaśniony" przez pozostałe zmienne znajdujące się w macierzy układu, (2) umieszczenie punktów reprezentujących kolumny na ostatecznym układzie współrzędnych dostarczy informacji o charakterze (kierunku) związku pomiędzy kolumnami oryginalnej macierzy układu a kolumnami określającymi fakt zachorowania. Ta technika wprowadzania punktów dodatkowych w wielowymiarowej analizie korespondencji jest niekiedy nazywana *odwzorowywaniem predykcyjnym* .

Tablica Burta.

Obliczenia występujące w przypadku wielowymiarowej analizy korespondencji są prowadzone nie na macierzy układu (która przy dużej liczbie obserwacji może mieć bardzo znaczne rozmiary), ale na iloczynie wewnętrznym tej macierzy, który nazywany jest macierzą *Burta*. Jeżeli dysponujemy tabelami częstości, to tablicę *Burta* otrzymamy dodając bloki

określające związki każdej zmiennej z sobą samą. Przykładowo tablica *Burta* dla prezentowanej poprzednio dwudzielczej tabeli liczności przedstawia się następująco:

	Grupa pracownicza					Intensywność palenia			
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
(1) Starsi kierownicy	11	0	0	0	0	4	2	3	2
(2) Młodszy kierownicy	0	18	0	0	0	4	3	7	4
(3) Starsi personel	0	0	51	0	0	25	10	12	4
(4) Młodszy personel	0	0	0	88	0	18	24	33	13
(5) Sekretarki	0	0	0	0	25	10	6	7	2
(1) Palenie:Nie pali	4	4	25	18	10	61	0	0	0
(2) Palenie:Mało	2	3	10	24	6	0	45	0	0
(3) Palenie:Średnio	3	7	12	33	7	0	0	62	0
(4) Palenie:Dużo	2	4	4	13	2	0	0	0	25

Tablica *Burta* ma jasno określoną strukturę. W przypadku dwóch zmiennych jakościowych (pokazanych powyżej) ma cztery części: (1) tabelę dwudzielczą zmiennej *Grupa pracownicza* z samą sobą, (2) tabelę dwudzielczą zmiennej *Grupa pracownicza* ze zmienną *Intensywność palenia*, (3) tabelę dwudzielczą zmiennej *Intensywność palenia* ze zmienną *Grupa pracownicza*, (4) tabelę dwudzielczą zmiennej *Intensywność palenia* z samą sobą. Zauważmy, że macierz ta jest symetryczna oraz że sumy elementów głównej przekątnej dotyczących związku danej zmiennej ze sobą muszą być zawsze takie same (ponieważ rozpatrujemy 193 obserwacje, dlatego suma elementów głównej przekątnej dotyczących zmiennej *Grupa pracownicza* wynosi 193 podobnie jak suma elementów głównej przekątnej w części dotyczącej *Intensywności palenia*).

Zauważmy, że elementy leżące poza główną przekątną w częściach opisujących związki zmiennej z samą sobą są równe *0*. W tych miejscach mogą pojawić się inne wartości, jeżeli tablica *Burta* została wyznaczona na podstawie macierzy układu przy stosowaniu kodowania rozmytego (patrz powyżej).

Tablica Burta

Tablica Burta jest iloczynem wewnętrznym macierzy układu, a wyniki wielowymiarowej analizy korespondencji są takie same, jak wyniki, jakie otrzymalibyśmy dla punktów reprezentujących kolumny przy zastosowaniu prostej analizy korespondencji do macierzy układu.

Załóżmy, że mamy dane dotyczące *przeżyć* osób w trzech grupach *wiekowych* z trzech *miejsc zamieszkania*, przedstawione w następującej tabeli:

	PRZEŻYCIE		WIEK			MIEJSCE ZAMIESZKANIA		
Nr obserwacji	**NIE**	**TAK**	**<50**	**50-69**	**69+**	**TOKYO**	**BOSTON**	**GLAMORGN**
1	0	1	0	1	0	0	0	1
2	1	0	1	0	0	1	0	0
3	0	1	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1

...	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.
762	1	0	0	1	0	1	0	0
763	0	1	1	0	0	0	1	0
764	0	1	0	1	0	0	0	1

W tym układzie danych, kod *1* został wprowadzony dla oznaczenia kategorii, do której (w danej zmiennej) należy konkretna obserwacja (na przykład zmienna *Przeżycie* ma warianty *Nie* oraz *Tak*). Na przykład osoba numer *1*, w wieku 50-69, pochodząca z *Glamorgn* nie zmarła w okresie objętym analizą. Przytoczony zbiór danych zawiera informacje o *764* osobach.

Jeżeli powyższą macierz układu (kodów) nazwiemy X, to iloczyn wewnętrzny macierzy X'X jest macierzą *Burta*. Dla omawianego przykładu przedstawiona jest ona poniżej.

	PRZEŻYCIE		**WIEK**			**MIEJSCE ZAMIESZKANIA**		
	NIE	**TAK**	**<50**	**50-69**	**69+**	**TOKYO**	**BOSTON**	**GLAMORGN**
PRZEŻYCIE:NIE	210	0	68	93	49	60	82	68
PRZEŻYCIE:TAK	0	554	212	258	84	230	171	153
WIEK:<50	68	212	280	0	0	151	58	71
WIEK:50-69	93	258	0	351	0	120	122	109
WIEK:69+	49	84	0	0	133	19	73	41
MIEJSCE ZAMIESZKANIA:TOKYO	60	230	151	120	19	290	0	0
MIEJSCE ZAMIESZKANIA:BOSTON	82	171	58	122	73	0	253	0
MIEJSCE ZAMIESZKANIA:GLAMORGN	68	153	71	109	41	0	0	221

Tablica *Burta* ma jasno określoną strukturę. Jest to macierz symetryczna. Przy trzech zmiennych jakościowych macierz ta składa się z 3 x 3 = 9 części utworzonych przez zestawienie w postaci tabeli wielodzielczej każdej ze zmiennych z nią samą (jak również kategorii wszystkich zmiennych). Zauważmy, że suma elementów głównej przekątnej leżących w trzech częściach zawierających główną przekątną jest stała (i w naszym przykładzie wynosi 764).

W naszym przykładzie, w tych częściach macierzy, które leżą wzdłuż głównej przekątnej, elementy poza przekątną są równe *0*. Elementy te mogą być różne od zera, jeżeli stosowano kodowanie rozmyte, wykorzystując na przykład prawdopodobieństwo do kodowania przynależności przypadków do grup.