Andrzej Tujaka

Zachodniopomorska Szkoła Biznesu w Szczecinie

Zastosowanie pakietu CAR do analizy korespondencji

Streszczenie

W artykule przedstawiono zastosowanie pakietu programów CAR, zrealizowanych w środowisku Matlab, do analizy danych tabelarycznych. Opisano budowę pakietu, dobór parametrów przetwarzania, podano podstawy teoretyczne metody analizy korespondencji oraz sposoby interpretacji wyników. Omówiono dwa tryby pracy pakietu, pod nadzorem przyjaznego sprzęgu użytkownika oraz za pomocą szeregu poleceń. Program CAR realizuje analizę korespondencji z zastosowaniem rotacji osi, zarówno ortogonalnych jak i ukośnych, umożliwiając uzyskanie prostej struktury danych. Analizę struktury danych ułatwia graficzne przedstawienie wyników za pomocą wykresów biplot.

Słowa kluczowe: analiza korespondencji, biplot, obroty ortogonalne, obroty ukośne, Matlab

Wstęp

Celem artykułu jest przedstawienie pakietu programów CAR[1][1] zrealizowanego w środowisku MATLAB[2][2], będącego implementacją analizy korespondencji. Pakiet posiada sprzęg użytkownika, umożliwiający wygodne wprowadzanie danych, parametrów przetwarzania oraz określanie postaci graficznej i numerycznej wyników. W części 2 opisano krótko cele i metody analizy korespondencji oraz formy przedstawiania graficznego wyników analizy w postaci wykresu zwanego biplotem. Opisano miary diagnostyczne służące do oceny wpływu różnych czynników na postać geometryczną wykresu. Dla uproszczenia interpretacji wyników zastosowano obroty ortogonalne i ukośne osi, prowadzące do tzw. prostej struktury wyników. W części 3 przedstawiono strukturę i sposób obsługi pakietu. W części 4 przedstawiono prosty przykład ilustrujący zastosowania pakietu do analizy danych. Część 5 zawiera krótkie podsumowanie doświadczeń wyniesionych z posługiwania się pakietem CAR oraz sformułowanie pewnych wniosków końcowych.

Analiza korespondencji

Wyczerpujące przedstawienie analizy korespondencji przedstawia M.J. Greenacre[3][3]. Analiza korespondencji jest jedną z metod redukcji wymiarowości danych, zazwyczaj

[1] U. Lorenco-Seva, M. van de Velden, H.A.L. Kiers, CAR: A MATLAB Package to Compute Correspondence Analysis with Rotations. Journal of Statistical Software, September 2009, Volume 31, Issue 8.

[2] The MathWorks Inc (2007). MATLAB – The Language of Technical Computing, Version 7.5. The MathWorks, Inc., Natick, Massachusetts. URL http:/www.mathworks.com/products/matlab/.

[3] M.J. Greenacre, Theory and Application of Correspondence Analysis. Academic Press, London, 1984, M.J. Greenacre, Correspondence Analysis in Practice. 2nd Ed. Chapman & Hall/CRC. London, 2007.

do dwóch lub trzech wymiarów, pozwalając na bezpośrednią analizę wzrokową danych i zależności między danymi. Analizę korespondencji stosuje się do danych nieujemnych, częstoprzedstawianych w postaci macierzy kontyngencji. Macierz kontyngencji przedstawia częstości występowania różnych kategorii dwóch danych kategorialnych: n kategorii jednej zmiennej stanowią wiersze, natomiast p kategorii drugiej zmiennej stanowią kolumny macierzy kontyngencji **F** o wymiarze n x p (macierzy klasyfikacji skrzyżowanej). Dążąc do interpretacji geometrycznej, wiersze macierzy **F** traktujemy jako n punktów w przestrzeni p wymiarowej, a kolumny jako p punktów w przestrzeni n wymiarowej.

Element f_{ij} macierzy **F** określa, ile razy zaobserwowano równoczesne wystąpienie kategorii i. pierwszej zmiennej oraz kategorii j. drugiej zmiennej. Na ogół liczbę tych wystąpień dzieli się przez całkowitą liczbę przeprowadzonych obserwacji. Macierz jest centrowana przy pomocy wektora **r** będącego sumą elementów każdego wiersza: $\mathbf{r} = \mathbf{F1}_n$, gdzie $\mathbf{1}_n$ jest wektorem kolumnowym zbudowanym z n jedynek, oraz wektora $\mathbf{c} = \mathbf{F}^T\mathbf{1}_p$, gdzie symbol ^{T} oznacza transponowanie. Tworzymy dwie macierze diagonalne $\mathbf{D}_r = \mathrm{diag}(\mathbf{r})$ oraz $\mathbf{D}_c = \mathrm{diag}(\mathbf{c})$, służące do normalizacji macierzy **F**: $\mathbf{F}_s = \mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{rc}^T)\mathbf{D}_c^{-1/2}$. Celem analizy korespondencji jest znalezienie macierzy współrzędnych **X** i **Y** o niskim wymiarze k = 2 lub 3, zawierających współrzędne punktów odpowiadającym wierszom oraz kolumnom macierzy $\mathbf{F}_s$. Macierze **X** i **Y** wybieramy tak, aby minimalizować funkcję kryterialną $f(\mathbf{X}, \mathbf{Y}) = \left\|\mathbf{F}s - \mathbf{D}_r^{1/2}\mathbf{XY}^T\mathbf{D}_c^{1/2}\right\|^2$, przy warunkach $\mathbf{X}^T\mathbf{D}_r\mathbf{X} = \mathbf{Y}^T\mathbf{D}_c\mathbf{Y} = \mathbf{I}$. Jeśli $\mathbf{F}_s = \mathbf{USV}^T$ jest rozkładem osobliwym macierzy $\mathbf{F}_s$, wówczas funkcja f(**X**, **Y**) jest minimalizowana przez: $\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}_k\mathbf{S}_k^a$ oraz $\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}_k\mathbf{S}_k^b$, gdzie $\mathbf{U}_k$ o wymiarze n x k, oraz $\mathbf{V}_k$ o wymiarze p x k są macierzami wektorów osobliwych odpowiadających k największym wartościom osobliwym, leżących na przekątnej macierzy $\mathbf{S}_k$. Parametry *a* i *b* określają rodzaj współrzędnych macierzy **X** i **Y**. Parametry *a* i *b* mogą mieć jedną z czterech kombinacji wartości:

1. *a* = 1 i *b* = 0, wówczas wiersze macierzy X nazywamy współrzędnymi głównymi, wiersze macierzy Y współrzędnymi standardowymi. Wykres łączny obu współrzędnych jest wykresem biplot. Odległości między punktami wierszy są (w przybliżeniu) odległościami chi-kwadrat. Ta kombinacja parametrów a i b nazywa się normalizacją główną wierszy.
2. *a* = 0 i *b* = 1, wówczas wiersze macierzy X nazywamy współrzędnymi standardowymi, wiersze macierzy Y są współrzędnymi głównymi. Wykres łączny obu współrzędnych jest biplotem. Odległości między punktami macierzy kolumn Y są (przybliżonymi) odległościami chi-kwadrat. Ta kombinacja parametrów a i b nazywa się normalizacją główną kolumn.
3. *a* = 0,5 i *b* = 0,5, wówczas współrzędne punktów wierszy X i współrzędne punktów kolumn Y nazywamy współrzędnymi symetrycznymi. Łączny wykres obu zbiorów punktów jest biplotem. Ten rodzaj wykresu nazywa się również biplotem kanonicznym.
4. *a* = 1 i *b* = 1, Łączny wykres punktów macierzy X i Y nazywa się francuskim wykresem symetrycznym. Odległości miedzy punktami macierzy wierszy X oraz miedzy punktami macierzy kolumn Y są (przybliżonymi) odległościami chi-kwadrat. Natomiast odległości miedzy punktami wierszy i kolumn są niezdefiniowane. Ponadto kąty między punktami wierszy i kolumn nie mają merytorycznej interpretacji.

W przypadkach 1 – 3 iloczyn $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}^T\mathbf{D}_c^{1/2}$ optymalnie aproksymuje macierz $\mathbf{F}_s$. Rozróżnienie współrzędnych głównych, standardowych i symetrycznych jest w analizie korespondencji bardzo ważne.

Współrzędne główne są współrzędnymi analizowanych zmiennych, i tak w przypadku $a = 1$ są to współrzędne związane ze zmiennymi wierszy, natomiast dla $a = 0$ ze zmiennymi kolumn.

Współrzędne standardowe są współrzędnymi zmiennych wspomagających opis analizowanych zmiennych; dla a = 0 współrzędne związane są ze zmiennymi wierszy *a* dla $a = 1$ ze zmiennymi kolumn.

W przypadku współrzędnych symetrycznych opis odnosi się zarówno do zmiennych wierszy jak i kolumn.

Współrzędne wierszy i kolumn są związane tzw. wzorami przejścia. Oznaczając Xs zbiór współrzędnych standardowych wierszy, współrzędne główne kolumn można otrzymać ze wzoru $\mathbf{Y}_p = \mathbf{D}_c^{-1}\mathbf{F}^T\mathbf{X}_s$, i analogicznie, współrzędne główne wierszy można otrzymać z zależności $\mathbf{X}_p = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{Y}_s$. Wzory te można wykorzystać do wykreślania dodatkowych wierszy lub kolumn na wykresie analizy korespondencji.

Wykreślając punkty odpowiadające wierszom i kolumnom wielowymiarowej macierzy kontyngencji tracimy część informacji. Ilość zachowanej informacji mierzymy stosunkiem sumy kwadratów k największych wartości osobliwych do sumy kwadratów wszystkich wartości osobliwych macierzy $\mathbf{F}_s$. Możemy również ocenić jakość współrzędnych wierszy i kolumn. Niech **X** oznacza k wymiarową macierz współrzędnych głównych wierszy. Dzieląc ważone kwadraty współrzędnych głównych przez inercję otrzymujemy tak zwane absolutne wkłady we współrzędne wierszy. Wkład absolutny i. wiersza w j. oś definiuje się jako $w_{ij} = (r_i/s_j^2)x_{ij}^2$. Termin absolutny odnosi się do wag r_i równych całkowitej liczbie obserwacji w wierszu, co ma znaczenie w obliczeniu wkładu punktów. Absolutny wkład ukazuje, jaki był udział współrzędnej w inercji opisanej w kierunku odpowiedniej osi. Wartości te są często używane do przypisania odpowiednich nazw k osiom użytym w aproksymacji. Względnie wysoki absolutny wkład określonego wiersza wskazuje na to iż ma on ważny wpływ na wyznaczenie położenia osi. Z tego powodu osie można nazywać stosownie do podzbioru zmiennych mających duży wkład. Ponadto uwzględnienie stosunku kwadratu współrzędnej głównej do sumy ważonej kwadratów współrzędnych wzdłuż k wymiarów uzyskuje się tak zwany udział względny w wierszach: udział względny j. osi w i. wierszu wynosi $s_{ij} = x_{ij}^2/\Sigma_{l=1}^{k}x_{il}^2$. Udziały względne są równe kwadratowi korelacji miedzy wierszem i osiami głównymi. Geometrycznie można je interpretować jako kwadraty kosinusów katów między każdym profilem wiersza i każdą osią główną. Udziały względne wskazują jak dobrze pewien punkt jest reprezentowany przez konkretną oś. Suma pierwszych k udziałów względnych informuje o jakości reprezentowania punktu w przestrzeni k wymiarowej.

Obroty ortogonalne i ukośne

Współrzędne wierszy i kolumn zapisane w macierzach **X** i **Y** są badane w celu wyjaśnienia znaczenia k współrzędnych. W wielu przypadkach rozwiązaniem najlepszym

jest rozwiązanie najłatwiejsze do interpretacji. W pracy[54] przedstawiono procedury obrotu ortogonalnego zapewniającego ułatwienie interpretacji osi współrzędnych, poprzez pomnożenie macierzy **X** i **Y** z prawej strony przez macierz obrotu T, co dalej zapewnia spełnienie warunku stawianego wykresom biplot, tj. by wyrażenie $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}^T\mathbf{D}_c^{1/2}$ optymalnie aproksymowało macierz $\mathbf{F}_s$. Wstawiając w to wyrażenie macierze współrzędnych otrzymanych w wyniku obrotu, **XT** i **YI**, ze względu na ortogonalność macierzy **T** ($\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}$), $\mathbf{D}_r^{1/2}\mathbf{X}\mathbf{T}(\mathbf{Y}\mathbf{T})^T\mathbf{D}_c^{1/2} = \mathbf{D}_r^{1/2}\mathbf{X}(\mathbf{T}\mathbf{T}^T)\mathbf{Y}^T\mathbf{D}_c^{1/2} = \mathbf{D}_r^{1/2}\mathbf{X}\mathbf{Y}^T\mathbf{D}_c^{1/2}$. W obrotach ukośnych prostsze jest wykonywanie obrotu albo macierzy **X** albo **Y**. Rozpatrzmy przypadek obracania tylko macierzy **Y** w celu uzyskania prostej struktury. Będziemy stosowali współrzędne standardowe dla **X** i szukali macierzy obrotu **U**, która maksymalizuje prostotę $\mathbf{Y}_o = \mathbf{Y}(\mathbf{U}^{T)-1}$, przy czym $\mathbf{X}_o = \mathbf{X}\mathbf{U}$. Nakładamy ograniczenie $\mathrm{diag}(\mathbf{U}^T\mathbf{U}) = \mathbf{I}$. Analogicznie, jeśli szukamy prostej struktury dla macierzy $\mathbf{X}_o$, przyjmujemy współrzędne standardowe dla **Y** i szukamy macierzy obrotu **U** maksymalizującej prostotę $\mathbf{X}_o = \mathbf{X}(\mathbf{U}^{T)-1}$, $\mathbf{Y}_o = \mathbf{Y}\mathbf{U}$ oraz $\mathrm{diag}(\mathbf{U}^T\mathbf{U}) = \mathbf{I}$.
Rozpatrzmy równoczesny obrót **X** i **Y** dla uzyskanie prostej struktury tych macierzy. Zazwyczaj wykonuje się to dla przypadku współrzędnych symetrycznych. Szukamy takiej macierzy obrotu **U** która pozwala uzyskać równocześnie prostą strukturę macierzy $\mathbf{X}_o = \mathbf{X}(\mathbf{U}^T)^{-1}$ i $\mathbf{Y}_o = \mathbf{Y}(\mathbf{U}^T)^{-1}$. W pracy[65] przedstawiono procedurę maksymalizacji prostoty przy ograniczeniu $\mathrm{diag}(\mathbf{U}^T\mathbf{U}) = \mathbf{I}$. Po wykonaniu obrotu macierze $\mathbf{X}_o$ i $\mathbf{Y}_o$ można wykorzystać do nazwania osi współrzędnych oraz określenia, które zmienne (wierszy i kolumn) są najbardziej związane z każda osią. W wymienionej pracy przyjęto obliczanie wartości średniej kwadratów współrzędnych dla każdej osi (wymiaru). Następnie porównuje się kwadrat każdej współrzędnej z odpowiadającej jej średniej, i tylko współrzędne których kwadrat jest większy od wartości średniej są wybrane jak liczące się współrzędne. Osie otrzymują nazwy zależnie od charakterystyk tak wybranych współrzędnych. Macierze **X** i **Y** są ważone przed wykonaniem obrotu za pomocą macierzy diagonalnych $\mathbf{W}_x$ i $\mathbf{W}_y$, tak iż obracane są macierze $\mathbf{W}_x\mathbf{X}$ i $\mathbf{W}_y\mathbf{Y}$. Macierze $\mathbf{W}_x\mathbf{X}$ i $\mathbf{W}_y\mathbf{Y}$ można wybrać na jeden z trzech sposobów: 1) macierze wag są macierzami jednostkowymi, czyli nie stosuje się wag, 2) stosujemy macierze wag $\mathbf{W}_x = \mathbf{W}_r^{1/2}$ oraz $\mathbf{W}_y = \mathbf{W}_c^{1/2}$. Taki sposób ważenia powoduje ulokowanie rzadko występujących punktów blisko początku układu współrzędnych, pozostawiając pozostałe punkty w dużej odległości od środka, 3) stosujemy macierze normujące wiersze obracanych macierzy, $\mathbf{W}_x = \mathrm{diag}(\mathbf{X}^T\mathbf{X})$ oraz $\mathbf{W}_y = \mathrm{diag}(\mathbf{Y}^T\mathbf{Y})$. Przy takim ważeniu wszystkie wiersze mają taki sam wpływ na końcowe położenie osi współrzędnych.

Biplot

Biplotem nazywamy przedstawienie graficzne prezentujące relacje miedzy dwoma zbiorami punktów. Jeden zbiór punktów reprezentuje np. obiekty, respondentów, itp., a drugi zbiór reprezentuje cechy, atrybuty, odpowiedzi na pytania testowe, itp. zmienne

[4] M. Van de Velden, H.A.L Kiers. Rotation in Correspondence Analysis. Journal of Classification, 22, 2005, s. 251–271.

[5] U. Lorenzo-Seva, M. Van de Velden, H.A.L. Kiers. Oblique rotation in Correspondence Analysis a Step Forward in the Search of the Simplest Interpretation. British Journal of Mathematical and Statistical Psychology, 62, 2009, s. 583–600.

Często obiekty odpowiadają wierszom a cechy kolumnom macierzy danych. Podobieństwa obiektów/zmiennych charakteryzują odległości między punktami, odpowiadającymi obiektom/zmiennym. Kąt miedzy punktami obu zbiorów a początkiem układu współrzędnych charakteryzuje ich korelacje. Biploty w połączeniu z macierzami wyników ułatwiają interpretację związków miedzy obiektami i zmiennymi.

Opis pakietu CAR

Pakiet CAR składa się z trzech modułów:

1. Car(): uruchamia sprzęg użytkownika który steruje innymi składnikami pakietu. Przy zastosowaniu tej funkcji korzystanie z pozostałych dwóch modułów jest niepotrzebne.
2. Canalysis(): moduł obliczający analizę korespondencji. Na wyjściu programu tworzona jest macierz strukturalna zawierająca wszystkie macierze związane z analizą. Macierz wyjściową można wydrukować przy pomocy funkcji PrintDescriptives() oraz PrintCoordinates(). Funkcje te drukują macierze odpowiadające wybranemu modelowi współrzędnych wyspecyfikowanego w funkcji Canalysis(). Ponadto reprezentację graficzną można otrzymać posługując się funkcją Map().
3. ComputeRotation(): realizuje ortogonalne i ukośne obroty osi. Wyjściem jest macierz strukturalna zawierające macierze związane z rodzajem obrotu wybranego w funkcji ComputeRotation().

Plik zawierający pakiet CAR musi znajdować się w bieżącym katalogu (current directory), który można ustawić w środowisku Matlab, lub wybrać poleceniem cd ze ścieżka go katalogu zawierającego car: np. cd C:\users\desktop\car.

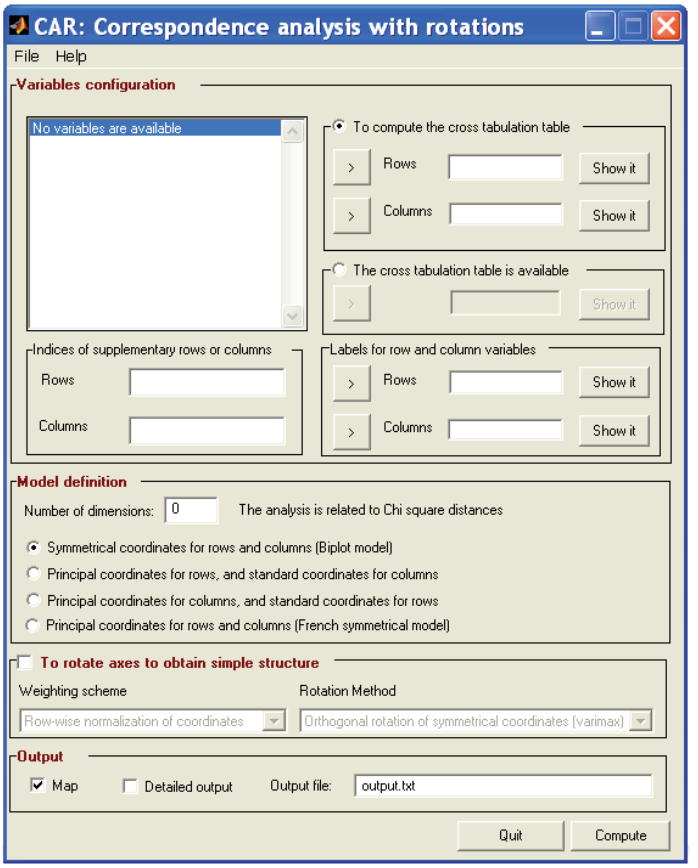
Sprzęg użytkownika uruchamia się wpisując w oknie poleceń Matlaba plecenie car. Wygląd sprzęgu pokazano na rysunku 1.

Aby wykonać analizę korespondencji posługując się sprzęgiem użytkownika, należy wykonać osiem kroków:

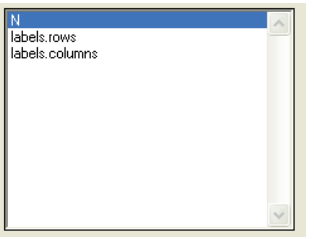
(1) jeśli analizowane dane są załadowane do przestrzeni roboczej, uruchamiamy sprzęg poleceniem car. Zmienne będą pokazywane w okienku zmiennych. Jeśli zmienne są strukturami, pokazane będą pola struktury. Na rysunku 2 pokazano przykładowe dane N oraz strukturą labels z dwoma polami (rows i columns).

Uwaga: wszystkie rysunki są kopiami ekranów generowanych w pakiecie CAR.

Rysunek 1. Wygląd sprzęgu użytkownika pakietu CAR

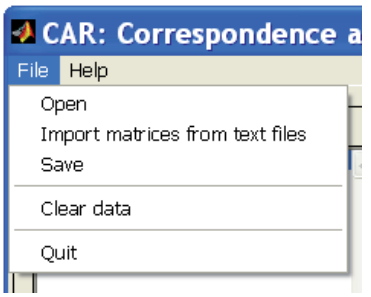


Rysunek 2. Wygląd okienka zmiennych



Jeśli dane nie są umieszczone w przestrzeni roboczej, można je załadować za pomocą opcji Open w menu File (patrz rys. 3)

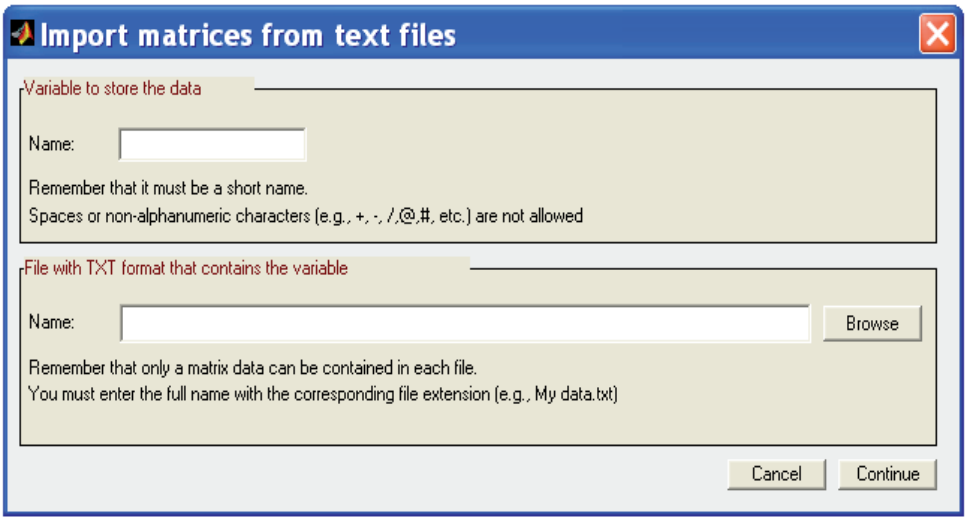
Rys.3. Opcje menu File



Źródło: Opracowanie własne

Do pamięci można również załadować dane zapisane w pliku ASCII, korzystając z opcji Import files from text files w menu File. Zostanie rozwinięte menu pokazane na rys. 4.

Rysunek 4. Opcja umożliwiająca ładowanie danych z plików tekstowych

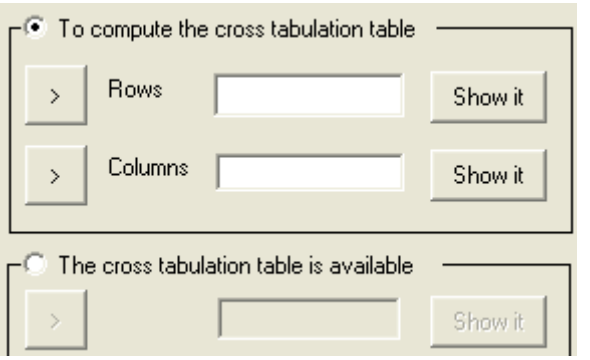


Źródło: Opracowanie własne

Dodatkowe informacje dotyczące danych wejściowych można uzyskać korzystając z menu Help z paska narzędzi menu głównego.

(2) Dane wejściowe muszą być zorganizowane w postaci macierzy kontyngencji. Jeżeli dysponujemy zmiennymi zawierającymi dane surowe, musimy wskazać zmienne które mają być traktowane jako wiersze oraz zmienne które mają stanowić kolumny macierzy kontyngencji ('cross tabulation table'). Zmienne muszą być zmiennymi numerycznymi. Natomiast jeśli dane są już zorganizowane w postaci macierzy, musimy wskazać nazwę macierzy w polu 'The cross tabulation table is available', patrz rys.5.

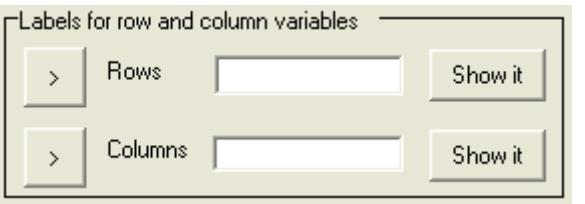
Rys.5. Menu organizowania danych do postaci macierzy kontyngencji lub wskazania nazwy macierzy kontyngencji ('cross tabulation table')



Źródło: Opracowanie własne

(3) Definiowanie etykiet. Jeśli dysponujemy danymi tekstowymi podającymi etykiety (nazwy) wierszy I kolumn macierzy kontyngencji, możemy wskazać ich nazwy posługując się pokazanym poniżej menu, patrz rys. 6.

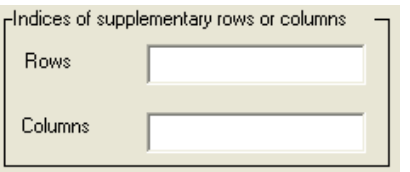
Rysunek 6. Menu do określania zmiennych zawierających etykiety wierszy i kolumn macierzy kontyngencji.



Źródło: Opracowanie własne

(4) Określanie dodatkowych wierszy i kolumn macierzy kontyngencji. Posługując się pokazanym poniżej menu (rys.7) można wskazać które wiersze lub kolumny macierzy kontyngencji należy traktować jako dodatkowe punkty w analizie korespondencji. Należy wpisać numery wierszy i kolumn macierzy kontyngencji. Można wskazywać więcej niż jeden wiersz lub kolumny które mają być traktowane jako dodatkowe. Numery wierszy (kolumn muszą być oddzielone spacją.

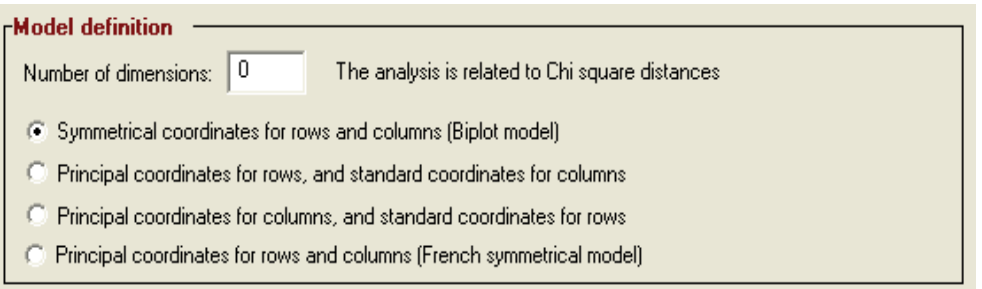
Rysunek 7. Wskazywanie numerów wierszy i kolumn traktowanych jako dodatkowe



Źródło: Opracowanie własne

(5) Definiowanie modelu osi. W menu pokazanym na rys. 8 wskazuje się liczbę wymiarów oraz rodzaje osi układu współrzędnych.

Rysunek 8. Okienko zadawania wymiarowości i rodzaju osi

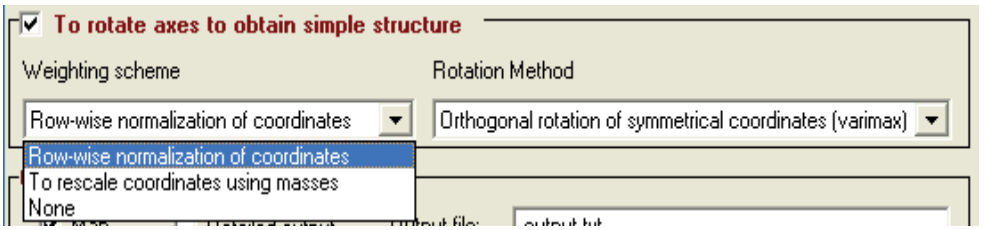


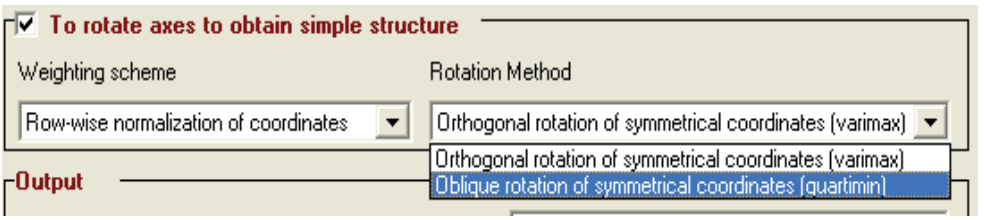
Źródło: Opracowanie własne

(6) Zadawanie obrotu osi

Do obliczenia obrotu osi należy zadać rodzaj zastosowanych wag oraz technikę obrotu. Dostępny rodzaj techniki zależy od rodzaju osi modelu. Parametry te zadaje się w okienkach przedstawionych na rysunku 9.

Rysunek 9. Widok okienek zadawania rodzaju wag i metody obrotu



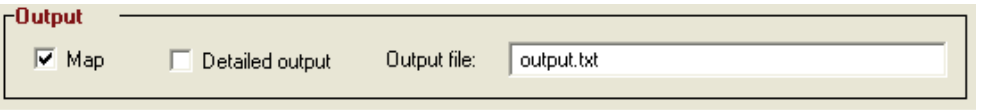


Źródło: Opracowanie własne

(7) Konfigurowanie opcji wyjścia

W okienku konfigurowania wyjścia określamy, czy ma być generowany wykres, poziom szczegółów danych wyjściowych oraz nazwę pliku danych wyjściowych. Wygląd okna określania tych informacji przedstawia rys. 10.

Rysunek 10. Okno określania czy ma być generowany wykres punktów wierszy i kolumn, poziom szczegółów danych wyjściowych i nazwa pliku wyjściowego



Źródło: Opracowanie własne

(8) Uruchomienie obliczeń

Po wybraniu parametrów analizy rozpoczynamy proces obliczeniowy wciskając przycisk Compute

Rysunek 11. Przycisk Compute

Źródło: Opracowanie własne

Macierze zawierające wyniki obliczeń są dostępne w strukturze 'output' i 'rotation', które będą dostępne do ewentualnych dalszych obliczeń po wydaniu poleceń:

output = getappdata(0, 'output');

rotation = getappdata(0, 'rotation');

Wszystkie obliczenia w ramach analizy korespondencji można wykonać bez posługiwania się sprzęgiem użytkownika. Najpierw należy przejść do katalogu zawierającego programy pakietu car oraz załadować do przestrzeni roboczej dane wejściowe (w przykładzie poniżej katalog car i dane N):

cd C:\users\desktop\car

load N;

Następnie szereg poniższych poleceń realizuje procedury analizy korespondencji:

output = canalysis(N,k,labels_x,labels_y);

rotation = ComputeRotation(output, k, coordinates, method, weights);

PrintDescriptives(output,detailed);

PrintCoordinates(output,k,coordinates,detailed);

PrintRotation(rotation,method, weights, output.labels, detailed);

map(1,2,0);

gdzie zastosowano oznaczenia:

N – macierz kontyngencji,

K – liczba wymiarów

labels_x i labels_y – etykiety wierszy i kolumn

coordinates – rodzaj współrzędnych zastosowanych w analizie, dostępne opcje:

1. Współrzędne symetryczne (model biplot)
2. Współrzędne główne wierszy i współrzędne standardowe kolumn

3. Współrzędne główne kolumn i współrzędne standardowe wierszy
4. Współrzędne główne wierszy I kolumn (model symetryczny francuski)

Method – metoda rotacji, dostępne opcje:

1. Obrót ortogonalny współrzędnych symetrycznych
2. Obrót ukośny współrzędnych symetrycznych
3. Obrót ortogonalny wspórzędnych głównych
4. Obrót ortogonalny współrzędnych standardowych
5. Obrót ukosny współrzędnych głównych

weights – sposób ważenia użyty w obrotach, dostępne opcje:

1. Normalizacja wierszowa współrzędnych
2. Skalowanie współrzędnych według mas
3. Bez skalowania.

Detailed – kontrola poziomu szczegółowości danych wyjściowych, dostępne opcje:

1. Wydruk wszystkich szczegółów
2. Wydruk podstawowych informacji

Jeśli chcemy dołaczyć dodatkowe wiersze i kolumny, wówczas polecenie

output = canalysis(N,k, labels_x,labels_y);

należy zastąpić poleceniem

output = canalysis(N,k, labels_x,labels_y, Sup, labels_sup);

gdzie:

Sup – struktura zawierająca macierze kontyngencji 'r' i/lub 'c' dodatkowych wierszy i kolumn

Labels – struktura zawierająca etykiety dodatkowych wierszy w łańcuchu 'r' i kolumn 'c'

Można również wykonać funkcję 'canalysis' z parametrem output = canalysis(N); wówczas liczba wymiarów wynosi 2 i zastosowane są domyślne etykiety wierszy i kolumn.

Przykłady analizy korespondencji

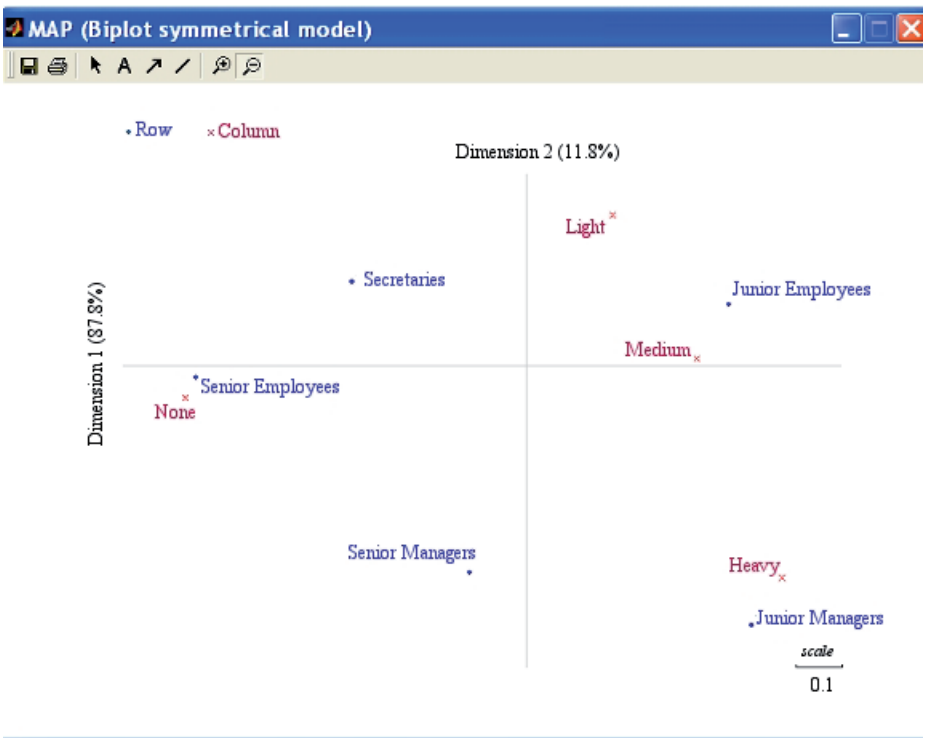
Rozpatrzmy prosty przykład analizy korespondencji danych dotyczący palących pracowników pewnej firmy (dane fikcyjne z pracy [3], str. 55). Palacze należą do jednej z kategorii: niepalący, palący mało, palący średnio, palący dużo, w firmie zajmują stanowiska: kierownicy wysokiego szczebla, kierownicy niskiego szczebla, pracownicy wykfalifikowani, pracownicy niewykfalikowani, pracownicy administracyjni. Macierz kontyngencji ma wymiar 4 x 5. Analizę korespondencji przeprowadzono dla 2 wymiarów, przyjmując model współrzędnych symetryczny, co umożliwia równoczesne przedstawienie klas palaczy i stopnia nałogu, oraz związki między nimi (a = 0,5 i b = 0,5). Ponadto dla uzyskania prostej struktury wykonano obrót ukośny. Przyjęto skalowanie współrzędnych według wag. Na rys. 12 pokazano wykres przed wykonaniem obrotu. Szczegółowe dane wyjściowe składa się ze 146 wierszy zapisanych w zbiorze tekstowym output.txt Dzięki obrotowi uzyskano bardzo prostą strukturę. Uzyskany wymiar d1 związany jest wskazując że Kierownicy wyższego stopnia mają tendencję do niepalenia, natomiast pracownicy niewykfalifikowani wykazują skłonność do palenia. Wymiar d1 jest dwubiegunowy, tj. oceny na tej osi są dodatnie i ujemne. Drugi

wymiar d2 jest wymiarem jednobiegunowym, pokazuje że kierownicy, szczególnie kierownicy niższego szczebla mają tendencję do palenia dużo. Korelacja wymiarów wynosi 0,25. Widać że wymiary są związane również z pozycją zawodową, i pracownicy każdego szczebla nie są związani tylko z natężeniem palenia. Najbardziej skomplikowaną grupą są pracownicy administracyjni: są wśród nich zarówno mało palący jak i niepalący. Kierownicy wyższego szczebla tworzą również skomplikowaną grupę: są zarówno palący dużo jak i jak i niepalący. Polecenie MAP wyświetla wykres we współrzędnych nieobróconych.

Zakończenie

Pakiet CAR umożliwia przeprowadzenie analizy korespondencji w przyjaznym dla użytkownika środowisku systemu Matlab, w szczególności oferując prosty sprzęg użytkownika. Wyniki analizy korespondencji umożliwiają przedstawienie graficzne danych wielowymiarowych Główną zaletą reprezentacji graficznej jest łatwość interpretacji i przekazywania złożonej informacji.

Rysunek 12. Dwuwymiarowy wykres korespondencji we współrzędnych symetrycznych



Źródło: Opracowanie własne

Analiza korespondencji umożliwia:

1. wyznaczenie położenia punktów reprezentujących obiekty (wiersze macierzy kontyngencji) oraz punktów reprezentujących ich atrybuty (kolumny macierzy kontyngencji) względem wspólnego układu odniesienia, umożliwiając opisanie zależności miedzy obiektami i ich atrybutami,
2. bezpośredni opis zależności miedzy obiektami i atrybutami, bez wprowadzania pośredniczących czynników,
3. łatwe dodawanie punktów odpowiadającym dodatkowym obiektom i atrybutom,
4. analizę bardzo dużych zbiorów danych wielowymiarowych.

Summary

Paper presents application of CAR package, implemented in Matlab environment, to analysis of contingency matrices. Structure of package as well as definition of processing parameters was presented, shortly presented the theoretical background of correspondence analysis and approaches to output interpretation. Two modes of CAR operation are possible: using the user friendly GUI or issuing commands in command lines. CAR implements a few rotation and axes scaling modes: accessible are orthogonal and oblique rotation, leading to simple data structure. Analysis of structure is simplified by graphical presentation as biplots.