Algorytmy eksploracji danych: Wykład 4

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Analiza korespondencji

- **Analiza korespondencji** (ang. *Correspondence Analysis (CA)*) jest techniką podobną do analizy składowych głównych (PCA) oraz analizy czynnikowej (FA). Dotyczy jednak wyłącznie **cech jakościowych**.

Analiza korespondencji

- **Analiza korespondencji** (ang. *Correspondence Analysis (CA)*) jest techniką podobną do analizy składowych głównych (PCA) oraz analizy czynnikowej (FA). Dotyczy jednak wyłącznie **cech jakościowych**.
- Cechy te mogą być wyrażone zarówno w **skali porządkowej** (np. mały, średni, duży itp.) jak i w **skali nominalnej** (np. biały, czerwony, zielony itp.).

Analiza korespondencji

- **Analiza korespondencji** (ang. *Correspondence Analysis (CA)*) jest techniką podobną do analizy składowych głównych (PCA) oraz analizy czynnikowej (FA). Dotyczy jednak wyłącznie **cech jakościowych**.
- Cechy te mogą być wyrażone zarówno w **skali porządkowej** (np. mały, średni, duży itp.) jak i w **skali nominalnej** (np. biały, czerwony, zielony itp.).
- Załóżmy na początku, że dysponujemy próbą losową n-elementową dwóch cech statystycznych o charakterze jakościowym: X i Y.

Tablica kontyngencji

Przeprowadzenie analizy korespondencji wymaga zgrupowania dostępnych danych w postaci tablicy, zwanej, z uwagi na jakościowy charakter zmiennych, **tablicą kontyngencji**

$$\mathbf{N} \stackrel{def}{=} \begin{bmatrix} n_{1,1} & n_{1,2} & \dots & n_{1,c} \\ n_{2,1} & n_{2,2} & \dots & n_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r,1} & n_{r,2} & \dots & n_{r,c} \end{bmatrix}$$

o r wierszach, odpowiadających wariantom (kategoriom) cechy X, oraz c kolumnach, odpowiadających wariantom (kategoriom) cechy Y. Element $n_{i,j}$ w tablicy kontyngencji oznacza liczbę elementów próby, dla których cecha X występuje w wariancie numer i, zaś cecha Y w wariancie numer j. Oczywiście $\sum_{i=1}^{r}\sum_{j=1}^{c} n_{i,j} = n$.

Miary współzależności cech jakościowych (1)

Podobnie jak analiza składowych głównych czy też analiza czynnikowa, również analiza korespondencji powinna być poprzedzona badaniem siły zależności pomiędzy obserwowanymi zmiennymi. Stosowane powszechnie **miary współzależności cech jakościowych** omówione są szczegółowo poniżej.

- **Statystyka** χ^2 jest definiowana w następujący sposób:

$$\chi^2 \stackrel{def}{=} \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j}-\widehat{n}_{i,j})^2}{\widehat{n}_{i,j}}, \tag{1}$$

gdzie $\widehat{n}_{i,j} \stackrel{def}{=} \frac{n_{i.}n_{.j}}{n}$ są tzw. **liczebnościami oczekiwanymi (teoretycznymi)**, przy czym $n_{i.} = \sum_{j=1}^{c} n_{i,j}$, $n_{.j} = \sum_{i=1}^{r} n_{i,j}$ są tzw. **liczebnościami brzegowymi** tablicy kontyngencji.

Miary współzależności cech jakościowych (2)

- Weźmy pod uwagę test hipotezy zerowej $H_0 : p_{i,j} = p_{i.}p_{.j}$ wobec alternatywy $H_1 : p_{i,j} \neq p_{i.}p_{.j}$, gdzie $p_{i,j}$, $p_{i.}$ oraz $p_{.j}$ oznaczają, odpowiednio, prawdopodobieństwo przyjęcia przez wektor (X, Y) wariantu (kategorii) i cechy X oraz wariantu (kategorii) j cechy Y, a także prawdopodobieństwa brzegowe: przyjęcia przez cechę X kategorii i oraz przez cechę Y kategorii j.

Miary współzależności cech jakościowych (2)

- Weźmy pod uwagę test hipotezy zerowej $H_0 : p_{i,j} = p_{i.}p_{.j}$ wobec alternatywy $H_1 : p_{i,j} \neq p_{i.}p_{.j}$, gdzie $p_{i,j}$, $p_{i.}$ oraz $p_{.j}$ oznaczają, odpowiednio, prawdopodobieństwo przyjęcia przez wektor (X, Y) wariantu (kategorii) i cechy X oraz wariantu (kategorii) j cechy Y, a także prawdopodobieństwa brzegowe: przyjęcia przez cechę X kategorii i oraz przez cechę Y kategorii j.

- Hipoteza zerowa oznacza niezależność cech X i Y. Można ją zweryfikować, wykorzystując tzw. **test niezależności** χ^2, który opiera się na statystyce testowej postaci (1).

Miary współzależności cech jakościowych (2)

- Weźmy pod uwagę test hipotezy zerowej $H_0 : p_{i,j} = p_{i.}p_{.j}$ wobec alternatywy $H_1 : p_{i,j} \neq p_{i.}p_{.j}$, gdzie $p_{i,j}$, $p_{i.}$ oraz $p_{.j}$ oznaczają, odpowiednio, prawdopodobieństwo przyjęcia przez wektor (X, Y) wariantu (kategorii) i cechy X oraz wariantu (kategorii) j cechy Y, a także prawdopodobieństwa brzegowe: przyjęcia przez cechę X kategorii i oraz przez cechę Y kategorii j.

- Hipoteza zerowa oznacza niezależność cech X i Y. Można ją zweryfikować, wykorzystując tzw. **test niezależności** χ^2, który opiera się na statystyce testowej postaci (1).
- Przy założeniu prawdziwości hipotezy H_0, statystyka ta ma rozkład chi kwadrat o $(r-1)(c-1)$ stopniach swobody, gdzie r i c oznaczają, odpowiednio, liczbę wariantów (kategorii) cechy X i cechy Y.

Miary współzależności cech jakościowych (3)

- Zbiór krytyczny testu ma postać $K = [\chi^2_{(r-1)(c-1),1-\alpha}, \infty)$, gdzie α jest ustalonym poziomem istotności testu.

Miary współzależności cech jakościowych (3)

- Zbiór krytyczny testu ma postać $K = [\chi^2_{(r-1)(c-1),1-\alpha}, \infty)$, gdzie α jest ustalonym poziomem istotności testu.
- W praktyce, im większa jest wartość statystyki χ^2 zdefiniowanej w (1), tym bardziej świadczy to o potencjalnej zależności cech X i Y.

Miary współzależności cech jakościowych (4)

- **Współczynnik V Craméra** zdefiniowany jest następująco:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}},$$

gdzie statystyka χ^2 została zdefiniowana w (1). Przyjmuje on wartości z przedziału $[0, 1]$. Im większa jego wartość, tym silniejszy związek pomiędzy badanymi zmiennymi. W praktyce przyjmuje się, że $V > 0.5$ świadczy o silnym związku pomiędzy X i Y.

Miary współzależności cech jakościowych (5)

- **Współczynnik C Pearsona** definiujemy w następujący sposób:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

gdzie n jest całkowitą licznością tablicy kontyngencji. Przyjmuje on wartości z przedziału $[0, 1]$, a jego interpretacja jest podobna do interpretacji współczynnika V Craméra.

Miary współzależności cech jakościowych (6)

- **Współczynnik T Czuprowa** definiujemy jako

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}.$$

Przyjmuje on wartości z przedziału $[0,1]$, a jego interpretacja jest podobna do współczynników V Craméra i C Pearsona.

Miary współzależności cech jakościowych (6)

- **Współczynnik T Czuprowa** definiujemy jako

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}.$$

Przyjmuje on wartości z przedziału $[0,1]$, a jego interpretacja jest podobna do współczynników V Craméra i C Pearsona.

- W przypadku porównywania siły zależności pomiędzy dwiema zmiennymi w dwóch różnych tablicach kontyngencji, tablice te powinny mieć jednakowe rozmiary (tę samą liczbę wierszy i kolumn).

Miary współzależności cech jakościowych (7)

Mamy jeszcze dwie inne miary, które znajdują zastosowanie w przypadku tablic dwudzielczych (dwa warianty każdej z badanych cech).

- **Współczynnik kontyngencji Yule'a** przyjmuje wartości z przedziału $[-1, 1]$ i jest zdefiniowany w następujący sposób:

$$Q = \frac{n_{1,1}n_{2,2} - n_{1,2}n_{2,1}}{n_{1,1}n_{2,2} + n_{1,2}n_{2,1}},$$

gdzie wielkości $n_{i,j}$ oznaczają odpowiednie liczebności dwudzielczej tablicy kontyngencji.

Miary współzależności cech jakościowych (7)

Mamy jeszcze dwie inne miary, które znajdują zastosowanie w przypadku tablic dwudzielczych (dwa warianty każdej z badanych cech).

- **Współczynnik kontyngencji Yule'a** przyjmuje wartości z przedziału $[-1, 1]$ i jest zdefiniowany w następujący sposób:

$$Q = \frac{n_{1,1}n_{2,2} - n_{1,2}n_{2,1}}{n_{1,1}n_{2,2} + n_{1,2}n_{2,1}},$$

gdzie wielkości $n_{i,j}$ oznaczają odpowiednie liczebności dwudzielczej tablicy kontyngencji.

- Jego interpretacja jest podobna do interpretacji klasycznego współczynnika korelacji liniowej Pearsona. Współczynnik kontyngencji Yule'a jest stosunkowo mało odporny na niewielkie liczebności obserwowane. Na przykład, w sytuacji, gdy $n_{i,j} = 0$ dla pewnej pary (i, j), może on wskazywać, wbrew prawdzie, na silną zależność pomiędzy badanymi zmiennymi.

Miary współzależności cech jakościowych (8)

- **Współczynnik kontyngencji** φ przyjmuje wartości z przedziału $[0, 1]$ i jest zdefiniowany następująco:

$$\varphi = \sqrt{\frac{\chi^2}{n}},$$

gdzie statystyka χ^2 została zdefiniowana w (1).

Macierz znaczników (1)

- Przed prezentacją samej techniki analizy korespondencji konieczne jest wprowadzenie całego szeregu pojęć i oznaczeń, które będą wykorzystywane w dalszej części wykładu.

Macierz znaczników (1)

- Przed prezentacją samej techniki analizy korespondencji konieczne jest wprowadzenie całego szeregu pojęć i oznaczeń, które będą wykorzystywane w dalszej części wykładu.
- Macierz Z_X wymiaru $n \times r$, opisująca próbę losową n-elementową ze względu na cechę X, w której każdym wierszu znajduje się tylko jedna jedynka, zaś pozostałe elementy wiersza są zerami, nazywamy **macierzą znaczników cechy** X.

Macierz znaczników (2)

Przykładowo, dla próby losowej 4-elementowej, w której cecha X występuje w 3 różnych wariantach, macierz Z_X może mieć postać

$$Z_X = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Zapis powyższy oznacza, że w pierwszym i trzecim elemencie próby cecha X występuje w trzecim wariancie (kategorii), w drugim elemencie - w pierwszej kategorii, zaś czwarty element próby charakteryzuje się występowaniem cechy X w drugim jej wariancie. Podobnie definiujemy macierz Z_Y wymiaru $n \times c$ jako **macierz znaczników cechy** Y. Zauważmy, że teraz $\mathbf{N} = Z_X^T \cdot Z_Y$.

Macierz korespondencji (1)

- **Macierzą korespondencji** nazywamy macierz $\mathbf{P} = (p_{i,j})$ wymiaru $r \times c$, której elementy są zdefiniowane jako $p_{i,j} = \frac{n_{i,j}}{n}$ (czyli są częstościami empirycznymi występowania konkretnych par wariantów cech X i Y).

Macierz korespondencji (1)

- **Macierzą korespondencji** nazywamy macierz $\mathbf{P} = (p_{i,j})$ wymiaru $r \times c$, której elementy są zdefiniowane jako $p_{i,j} = \frac{n_{i,j}}{n}$ (czyli są częstościami empirycznymi występowania konkretnych par wariantów cech X i Y).
- Częstości brzegowe macierzy korespondencji

$$p_{i.} = \sum_{j=1}^{c} p_{i,j} = \frac{n_{i.}}{n}$$

oraz

$$p_{.j} = \sum_{i=1}^{r} p_{i,j} = \frac{n_{.j}}{n}$$

nazywamy **masami wierszowymi** oraz **masami kolumnowymi** odpowiednio.

Macierz korespondencji (2)

Diagonalne macierze częstości wierszowych i kolumnowych definiujemy w następujący sposób:

$$\mathbf{D_r} \stackrel{def}{=} \begin{bmatrix} p_{1.} & 0 & \dots & 0 \\ 0 & p_{2.} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{r.} \end{bmatrix},$$

$$\mathbf{D_c} \stackrel{def}{=} \begin{bmatrix} p_{.1} & 0 & \dots & 0 \\ 0 & p_{.2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{.c} \end{bmatrix}. \qquad (2)$$

Macierz profili wierszy tablicy kontyngencji $\mathbf{N}$ określamy jako

$$\mathbf{R} \stackrel{def}{=} \mathbf{D_r}^{-1}\mathbf{P} = \begin{bmatrix} \frac{1}{p_{1.}} & 0 & \cdots & 0 \\ 0 & \frac{1}{p_{2.}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{p_{r.}} \end{bmatrix} \cdot \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,c} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r,1} & p_{r,2} & \cdots & p_{r,c} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{p_{1,1}}{p_{1.}} & \frac{p_{1,2}}{p_{1.}} & \cdots & \frac{p_{1,c}}{p_{1.}} \\ \frac{p_{2,1}}{p_{2.}} & \frac{p_{2,2}}{p_{2.}} & \cdots & \frac{p_{2,c}}{p_{2.}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{r,1}}{p_{r.}} & \frac{p_{r,2}}{p_{r.}} & \cdots & \frac{p_{r,c}}{p_{r.}} \end{bmatrix} = \left(\frac{p_{i,j}}{p_{i.}}\right) = \left(\frac{n_{i,j}}{n_{i.}}\right).$$

Macierz profili kolumn

Podobnie definiujemy **macierz profili kolumn** macierzy kontyngencji $\mathbf{N}$:

$$\mathbf{C} \stackrel{def}{=} \mathbf{P}\mathbf{D}_{\mathbf{c}}{}^{-1} = \begin{bmatrix} \frac{p_{1,1}}{p_{.1}} & \frac{p_{1,2}}{p_{.2}} & \cdots & \frac{p_{1,c}}{p_{.c}} \\ \frac{p_{2,1}}{p_{.1}} & \frac{p_{2,2}}{p_{.2}} & \cdots & \frac{p_{2,c}}{p_{.c}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{r,1}}{p_{.1}} & \frac{p_{r,2}}{p_{.2}} & \cdots & \frac{p_{r,c}}{p_{.c}} \end{bmatrix}$$

$$= \left(\frac{p_{i,j}}{p_{.j}} \right) = \left(\frac{n_{i,j}}{n_{.j}} \right).$$

Centrum wierszowe i kolumnowe (1)

- Wprowadza się także wektory kolumnowe $\mathbf{c}$ oraz $\mathbf{r}$, będące swego rodzaju **środkami ciężkości**, odpowiednio wierszy i kolumn (określa się je także **centrum wierszowym** oraz **centrum kolumnowym**).

Centrum wierszowe i kolumnowe (1)

- Wprowadza się także wektory kolumnowe $\mathbf{c}$ oraz $\mathbf{r}$, będące swego rodzaju **środkami ciężkości**, odpowiednio wierszy i kolumn (określa się je także **centrum wierszowym** oraz **centrum kolumnowym**).
- Centrum wierszowe to wektor kolumnowy postaci $\mathbf{c} \stackrel{def}{=} [p_{.1}, ..., p_{.c}]^T$, zaś centrum kolumnowe ma postać $\mathbf{r} \stackrel{def}{=} [p_{1.}, ..., p_{r.}]^T$.

Centrum wierszowe i kolumnowe (1)

- Wprowadza się także wektory kolumnowe $\mathbf{c}$ oraz $\mathbf{r}$, będące swego rodzaju **środkami ciężkości**, odpowiednio wierszy i kolumn (określa się je także **centrum wierszowym** oraz **centrum kolumnowym**).
- Centrum wierszowe to wektor kolumnowy postaci $\mathbf{c} \stackrel{def}{=} [p_{.1}, ..., p_{.c}]^T$, zaś centrum kolumnowe ma postać $\mathbf{r} \stackrel{def}{=} [p_{1.}, ..., p_{r.}]^T$.

- Oznaczając $\mathbf{1}_k \stackrel{def}{=} \underbrace{[1, 1, ..., 1]^T}_{k}$, możemy zauważyć, że macierze $\mathbf{R} - \mathbf{1}_r\mathbf{c}^T$ oraz $\mathbf{C} - \mathbf{r}\mathbf{1}_c^T$ (obie są macierzami wymiaru $r \times c$) są **miarami stopnia „odchylenia" od niezależności cech, odpowiednio wierszy i kolumn macierzy kontyngecji**.

Centrum wierszowe i kolumnowe (2)

- Rzeczywiście, elementy wiersza numer i pierwszej z tych macierzy mają postać $\frac{p_{i,j}}{p_{i.}} - p_{.j}$, czyli wskazują na stopień „odchylenia" profilu tego wiersza od średniego profilu wierszowego.

Centrum wierszowe i kolumnowe (2)

- Rzeczywiście, elementy wiersza numer i pierwszej z tych macierzy mają postać $\frac{p_{i,j}}{p_{i.}} - p_{.j}$, czyli wskazują na stopień „odchylenia” profilu tego wiersza od średniego profilu wierszowego.

- Podobnie, elementy kolumny numer j drugiej macierzy mają postać $\frac{p_{i,j}}{p_{.j}} - p_{i.}$, zatem wskazują, jak bardzo „odchyla” się profil tej kolumny od średniego profilu kolumnowego.

Rozkład SVD (1)

- Głównym celem analizy korespondencji jest przedstawienie analizowanego zbioru punktów w nowej przestrzeni, zwanej **przestrzenią rzutowania**, maksymalnie trójwymiarowej, przy zachowaniu niemal pełnej informacji, którą niesie oryginalna tablica kontyngencji.

Rozkład SVD (1)

- Głównym celem analizy korespondencji jest przedstawienie analizowanego zbioru punktów w nowej przestrzeni, zwanej **przestrzenią rzutowania**, maksymalnie trójwymiarowej, przy zachowaniu niemal pełnej informacji, którą niesie oryginalna tablica kontyngencji.
- W rozwiązaniu tego problemu wykorzystuje się tzw. **rozkład macierzy według wartości osobliwych** (ang. *Singular Value Decomposition (SVD)*).

Rozkład SVD (2)

- Weźmy pod uwagę dowolną niezerową macierz A wymiaru $n \times m$ (niekoniecznie kwadratową).

Rozkład SVD (2)

- Weźmy pod uwagę dowolną niezerową macierz A wymiaru $n \times m$ (niekoniecznie kwadratową).
- Metoda SVD polega na przedstawieniu tej macierzy w postaci

$$A = U \cdot \Gamma \cdot V^T,$$

gdzie Γ jest pewną macierzą diagonalną, przyjmijmy, że ma ona wymiar $k \times k$, na której głównej przekątnej znajdują się tzw. **wartości osobliwe** γ_i, $i = 1, ..., k$, macierzy A ułożone w porządku nierosnącym.

Rozkład SVD (2)

- Weźmy pod uwagę dowolną niezerową macierz A wymiaru $n \times m$ (niekoniecznie kwadratową).
- Metoda SVD polega na przedstawieniu tej macierzy w postaci

$$A = U \cdot \Gamma \cdot V^T,$$

gdzie Γ jest pewną macierzą diagonalną, przyjmijmy, że ma ona wymiar $k \times k$, na której głównej przekątnej znajdują się tzw. **wartości osobliwe** γ_i, $i = 1, ..., k$, macierzy A ułożone w porządku nierosnącym.

- Macierze: U - wymiaru $n \times k$ oraz V - wymiaru $m \times k$, są macierzami tzw. **lewych** oraz **prawych wektorów osobliwych**, odpowiednio.

Rozkład SVD (2)

- Weźmy pod uwagę dowolną niezerową macierz A wymiaru $n \times m$ (niekoniecznie kwadratową).
- Metoda SVD polega na przedstawieniu tej macierzy w postaci

$$A = U \cdot \Gamma \cdot V^T,$$

gdzie Γ jest pewną macierzą diagonalną, przyjmijmy, że ma ona wymiar $k \times k$, na której głównej przekątnej znajdują się tzw. **wartości osobliwe** γ_i, $i = 1, ..., k$, macierzy A ułożone w porządku nierosnącym.

- Macierze: U - wymiaru $n \times k$ oraz V - wymiaru $m \times k$, są macierzami tzw. **lewych** oraz **prawych wektorów osobliwych**, odpowiednio.
- Kolumny macierzy U są wektorami własnymi macierzy $A^T A$, natomiast kolumny macierzy V - wektorami własnymi macierzy AA^T.

Rozkład SVD (3)

- Wartości osobliwe macierzy A są pierwiastkami arytmetycznymi wartości własnych λ_k macierzy $A^T A$ oraz AA^T (wartości własne tych macierzy są identyczne).

Rozkład SVD (3)

- Wartości osobliwe macierzy A są pierwiastkami arytmetycznymi wartości własnych λ_k macierzy $A^T A$ oraz AA^T (wartości własne tych macierzy są identyczne).
- Mamy zatem następujące równości:

$$\begin{cases} A^T A = U \Lambda U^T, \\ A A^T = V \Lambda V^T, \end{cases}$$

gdzie $\Lambda = \Gamma^2$.

Rozkład SVD (3)

- Wartości osobliwe macierzy A są pierwiastkami arytmetycznymi wartości własnych λ_k macierzy $A^T A$ oraz AA^T (wartości własne tych macierzy są identyczne).
- Mamy zatem następujące równości:

$$\begin{cases} A^T A = U \Lambda U^T, \\ AA^T = V \Lambda V^T, \end{cases}$$

gdzie $\Lambda = \Gamma^2$.

- W analizie korespondencji rozkład według wartości osobliwych przeprowadza się dla tzw. **macierzy różnic standaryzowanych** (ang. *standarized residuals*), czyli ważonych odchyleń profili od centrum wierszowego i kolumnowego.

Rozkład SVD (3)

- Wartości osobliwe macierzy A są pierwiastkami arytmetycznymi wartości własnych λ_k macierzy $A^T A$ oraz AA^T (wartości własne tych macierzy są identyczne).
- Mamy zatem następujące równości:

$$\begin{cases} A^T A = U \Lambda U^T, \\ A A^T = V \Lambda V^T, \end{cases}$$

gdzie $\Lambda = \Gamma^2$.

- W analizie korespondencji rozkład według wartości osobliwych przeprowadza się dla tzw. **macierzy różnic standaryzowanych** (ang. *standarized residuals*), czyli ważonych odchyleń profili od centrum wierszowego i kolumnowego.
- Macierz ta ma następującą postać:

$$A = \mathbf{D_r}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D_c}^{-1/2}. \qquad (3)$$

Nowe współrzędne

Nowe współrzędne kategorii obydwu badanych cech są obliczane w następujący sposób:

- dla wierszy (kategorii cechy X) są to kolejne kolumny macierzy $F = \mathbf{D_r}^{-1/2} U \Gamma$;
- dla kolumn (kategorii cechy Y) są to kolejne kolumny macierzy $G = \mathbf{D_c}^{-1/2} V \Gamma$.

Inercja całkowita

- Udowodnić można ważny związek pomiędzy statystyką χ^2, obliczoną dla danej macierzy kontyngencji $\mathbf{N}$, a wartościami osobliwymi macierzy A różnic standaryzowanych określonej w (3):

$$tr(A^T A) = tr(AA^T) = tr(\Lambda) = \frac{\chi^2}{n} = \lambda = \sum_{k=1}^{rz(A)} \gamma_k^2,$$

gdzie $rz(A)$ oznacza rząd macierzy A, natomiast

$$\lambda = \sum_{k=1}^{rz(A)} \lambda_k = \sum_{k=1}^{rz(A)} \gamma_k^2 \qquad (4)$$

jest tzw. **inercją całkowitą**, która jest interpretowana jako miara stopnia zależności pomiędzy badanymi cechami.

Inercja całkowita

- Udowodnić można ważny związek pomiędzy statystyką χ^2, obliczoną dla danej macierzy kontyngencji $\mathbf{N}$, a wartościami osobliwymi macierzy A różnic standaryzowanych określonej w (3):

$$tr(A^T A) = tr(AA^T) = tr(\Lambda) = \frac{\chi^2}{n} = \lambda = \sum_{k=1}^{rz(A)} \gamma_k^2,$$

gdzie $rz(A)$ oznacza rząd macierzy A, natomiast

$$\lambda = \sum_{k=1}^{rz(A)} \lambda_k = \sum_{k=1}^{rz(A)} \gamma_k^2 \qquad (4)$$

jest tzw. **inercją całkowitą**, która jest interpretowana jako miara stopnia zależności pomiędzy badanymi cechami.

- Inercja całkowita, zwana także **całkowitą bezwładnością**, odpowiada całkowitej wariancji wyjściowego układu zmiennych w analizie składowych głównych i analizie czynnikowej.

Inercje główne (1)

- Pierwsze kolumny macierzy F i G zawierają, odpowiednio, współrzędne wierszy i kolumn (czyli kategorii cech X i Y) na pierwszej osi (osi głównej) układu współrzędnych.

Inercje główne (1)

- Pierwsze kolumny macierzy F i G zawierają, odpowiednio, współrzędne wierszy i kolumn (czyli kategorii cech X i Y) na pierwszej osi (osi głównej) układu współrzędnych.
- Jeżeli powiązania pomiędzy poszczególnymi kategoriami cech X i Y decydujemy się przedstawić w przestrzeni jednowymiarowej, poprzestajemy na pierwszych kolumnach macierzy F i G.

Inercje główne (1)

- Pierwsze kolumny macierzy F i G zawierają, odpowiednio, współrzędne wierszy i kolumn (czyli kategorii cech X i Y) na pierwszej osi (osi głównej) układu współrzędnych.
- Jeżeli powiązania pomiędzy poszczególnymi kategoriami cech X i Y decydujemy się przedstawić w przestrzeni jednowymiarowej, poprzestajemy na pierwszych kolumnach macierzy F i G.

- Prezentując wyniki w przestrzeni dwuwymiarowej, wykorzystujemy dwie pierwsze kolumny macierzy F i G (dwie pierwsze osie główne).

Inercje główne (1)

- Pierwsze kolumny macierzy F i G zawierają, odpowiednio, współrzędne wierszy i kolumn (czyli kategorii cech X i Y) na pierwszej osi (osi głównej) układu współrzędnych.
- Jeżeli powiązania pomiędzy poszczególnymi kategoriami cech X i Y decydujemy się przedstawić w przestrzeni jednowymiarowej, poprzestajemy na pierwszych kolumnach macierzy F i G.

- Prezentując wyniki w przestrzeni dwuwymiarowej, wykorzystujemy dwie pierwsze kolumny macierzy F i G (dwie pierwsze osie główne).
- Poszczególne wartości λ_k nazywamy **inercjami głównymi**. Odpowiadają one kolejnym osiom głównym układu współrzędnych. Inercja całkowita λ jest zatem sumą inercji głównych (porównaj (4)).

Inercje główne (2)

Jeżeli wartość inercji całkowitej jest niewielka, to powiązania pomiędzy kategoriami cech są słabe. W efekcie punkty odpowiadające poszczególnym kategoriom są skupione wokół centrum rzutowania (początku układu), którym jest środek ciężkości wierszy i kolumn.