

Artur Zaborski

WYKORZYSTANIE ANALIZY PROCRUSTESA W PRZYPADKU WIELU WYNIKÓW SKALOWANIA WIELOWYMIAROWEGO

1. Wstęp

Skalowanie wielowymiarowe jest zbiorem technik graficznej prezentacji zależności między obiektami traktowanymi jako punkty w przestrzeni wielowymiarowej. Na podstawie macierzy odległości między obiektami, wykorzystując odpowiednie procedury zależne od typu zmiennych, dokonuje się takiego rozmieszczenia na mapie percepcyjnej punktów, aby dopasowanie konfiguracji odległości w przestrzeni wielowymiarowej i dwuwymiarowej było jak najlepsze.

Takie przekształcenia, jak jednokładność, translacja czy rotacja układu punktów nie powodują zmiany proporcji odległości między nimi, dlatego dla danego zbioru obiektów w wyniku skalowania wielowymiarowego można otrzymać różne konfiguracje końcowe.

Często trzeba porównać konfigurację punktów reprezentujących badane obiekty z konfiguracją punktów reprezentującą te same obiekty, ale otrzymaną w wyniku innego badania, na podstawie innych danych źródłowych (np. pochodzących od różnych respondentów) czy też z wykorzystaniem innego programu komputerowego. Metoda dopasowująca jedną konfigurację do drugiej (tzw. konfiguracji wzorcowej) oraz wskazująca miarę tego dopasowania nosi nazwę analizy Procrustesa.

W artykule zaprezentowano uogólnioną analizę Procrustesa, wykorzystywaną do dopasowywania wielu wyników skalowania wielowymiarowego.

2. Zasady analizy Procrustesa

Przyjmijmy, że dane są dwie konfiguracje n punktów w euklidesowej przestrzeni p -wymiarowej uzyskane w wyniku skalowania wielowymiarowego, których

współrzędne tworzą macierze $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$ i $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]^T$ o wymiarach $n \times p$. Punkty obydwu konfiguracji reprezentują te same obiekty lub zmienne, co oznacza, że punkt $\mathbf{x}_i$ odpowiada punktowi $\mathbf{y}_i$ dla każdego $i = 1, 2, \dots, n$. Dopasowanie obydwu konfiguracji jest najlepsze jeżeli suma kwadratów odległości między odpowiadającymi sobie punktami $\mathbf{X}$ i $\mathbf{Y}$ jest najmniejsza, czyli gdy funkcja dopasowania przyjmująca postać

$$R^2 = \text{tr}(\mathbf{X} - \mathbf{Y})^T (\mathbf{X} - \mathbf{Y}) \quad (1)$$

osiąga wartość najmniejszą.

W celu dopasowania do konfiguracji $\mathbf{Y}$, zbiór punktów $\mathbf{X}$ poddaje się przekształceniom, którymi są translacja, rotacja i jednokładność.

Współrzędne $\mathbf{X}$ po wymienionych przekształceniach wynoszą:

$$\mathbf{x}'_i = \varphi \mathbf{A}^T \mathbf{x}_i + \mathbf{b}, \quad (2)$$

gdzie $\mathbf{A}$ oznacza ortogonalną macierz rotacji, $\mathbf{b}$ – wektor translacji, φ – stosunek jednokładności.

W nieważonej analizie Procrustes dopasowanie konfiguracji $\mathbf{X}$ do konfiguracji $\mathbf{Y}$, przy minimalizacji funkcji dopasowania (1), przebiega przez następujące etapy (zob. [5]):

1. Translacja $\mathbf{X}$ i $\mathbf{Y}$ tak, aby ich środki ciężkości znajdowały się w początku układu współrzędnych.

2. Wymnożenie macierzy $\mathbf{X}$ przez macierz rotacji $\mathbf{A} = (\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X})^{1/2} (\mathbf{Y}^T \mathbf{X})^{-1}$.

3. Znalezienie obrazu $\mathbf{X}$ przy jednokładności o stosunku $\hat{\varphi} = \frac{\text{tr}(\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X})^{1/2}}{\text{tr}(\mathbf{X}^T \mathbf{X})}$.

4. Obliczenie miary dopasowania $R^2 = 1 - \frac{\left\{ \text{tr}(\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X})^{1/2} \right\}^2}{\text{tr}(\mathbf{X}^T \mathbf{X}) \text{tr}(\mathbf{Y}^T \mathbf{Y})}$.

W analizie Procrustes etap pierwszy jest pomijany, gdy środki ciężkości $\mathbf{X}$ i $\mathbf{Y}$ są w początku układu współrzędnych, zaś etap trzeci – gdy suma kwadratów odległości punktów od początku układu współrzędnych jest dla obydwu konfiguracji taka sama¹ (zob. [1, s. 362]). W takim wypadku funkcję dopasowania można zapisać jako:

¹ W dalszej części artykułu przyjmuje się te założenia, ponieważ większość algorytmów skalowania wielowymiarowego dokonuje takiego rozmieszczenia punktów na mapie percepcyjnej, aby środek ciężkości punktów znajdował się w początku układu współrzędnych, a średnia kwadratów odległości punktów od początku układu współrzędnych wynosiła jeden.

$$R^2 = \text{tr}(\mathbf{Y} - \mathbf{XA})^T (\mathbf{Y} - \mathbf{XA}). \quad (3)$$

Jeżeli udział poszczególnych punktów reprezentujących obiekty lub zmienne w R^2 określają wagi o wartościach w_i ($i = 1, 2, \dots, m$), to zadaniem analizy Procrustes jest znalezienie takiej ortogonalnej macierzy rotacji $\mathbf{A}$, która minimalizuje wartość funkcji dopasowania

$$R^2 = \text{tr}(\mathbf{Y} - \mathbf{XA})^T \mathbf{W}_n^2 (\mathbf{Y} - \mathbf{XA}), \quad (4)$$

gdzie $\mathbf{W}_n = \text{diag}(w_1, w_2, \dots, w_n)$.

Funkcję dopasowania (4) możemy zapisać jako

$$R^2 = \text{tr}(\mathbf{W}_n \mathbf{Y} - \mathbf{W}_n \mathbf{XA})^T (\mathbf{W}_n \mathbf{Y} - \mathbf{W}_n \mathbf{XA}), \quad (5)$$

zatem macierz $\mathbf{A}$ wyznaczana jest tak jak dla modelu nieważonego, przy czym w miejsce macierzy $\mathbf{X}$ i $\mathbf{Y}$ pojawiają się odpowiednio $\mathbf{W}_n \mathbf{X}$ i $\mathbf{W}_n \mathbf{Y}$.

W przypadku gdy ważne są nie punkty, ale poszczególne wymiary macierzy $\mathbf{Y}$, funkcja dopasowania przyjmuje postać²:

$$R^2 = \text{tr}(\mathbf{Y} - \mathbf{XA})^T \mathbf{W}_p^2 (\mathbf{Y} - \mathbf{XA}), \quad (6)$$

gdzie $\mathbf{W}_p = \text{diag}(w_1, w_2, \dots, w_p)$.

3. Uogólniona analiza Procrustes

Załóżmy, że dysponujemy m macierzami niepodobieństw Δ_k ($k = 1, 2, \dots, m$), przy czym każda z nich przedstawia niepodobieństwa między tymi samymi obiektami postrzegane przez różnych respondentów. Na podstawie macierzy Δ_k wyznacza się k konfiguracji punktów $\mathbf{X}_k$ reprezentujących obiekty. Otrzymujemy w ten sposób trójwymiarową kostkę zmiennych o współrzędnych x_{iak} , gdzie:

$i = 1, 2, \dots, n$ oznacza liczbę badanych obiektów,

$a = 1, 2, \dots, p$ – wymiar przestrzeni, w której prezentowane są wyniki skalowania wielowymiarowego,

$k = 1, 2, \dots, m$ – liczba konfiguracji punktów.

Wzrokowe porównanie $\mathbf{X}_k$ może być utrudnione, ponieważ różnice w układach konfiguracji punktów na mapach percepcyjnych nie zawsze wynikają z różnic w postrzeganiu obiektów przez poszczególnych respondentów [5]. Dlatego porównanie wyników skalowania wielowymiarowego powinno być poprzedzone analizą Procrustes.

² Sposoby wyznaczenia macierzy $\mathbf{A}$ minimalizującej wartość (6) prezentują m.in. Cox i Cox [3, s. 133-134].

Wykorzystując uogólnioną analizę Procrustes, dokonuje się transformacji indywidualnych konfiguracji $\mathbf{X}_k$, tak aby każda z nich została jak najlepiej dopasowana do pozostałych. Podobnie jak w analizie dla dwóch wyników skalowania wielowymiarowego, dopuszczalnymi przekształceniami są translacja, rotacja i jednokładność, czyli te, które zachowują proporcje odległości między punktami.

Uogólniona analiza Procrustes prowadzi do wyznaczenia takich konfiguracji $\tilde{\mathbf{X}}_k$, dla których funkcja dopasowania (por. [1, s. 362])

$$R^2 = \sum_{k < l}^m \text{tr}(\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l)^T (\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l), \quad (7)$$

gdzie $\tilde{\mathbf{X}}_k = \varphi_k \mathbf{X}_k \mathbf{A}_k^T + \mathbf{1} \mathbf{b}_k^T$ ($k, l = 1, 2, \dots, m$), przyjmuje wartość najmniejszą.

Algorytm minimalizujący wartość funkcji R^2 ma charakter iteracyjny i polega na cyklicznym dopasowywaniu kolejnych konfiguracji punktów (tak jak w przypadku analizy Procrustes dla dwóch wyników skalowania) przy założeniu, że pozostałe są ustalone.

W tym celu wprowadza się średnią ze wszystkich $\tilde{\mathbf{X}}_k$ konfigurację (*centroid configuration*) $\mathbf{Z} = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{X}}_k$. Wtedy funkcja (7) przyjmuje postać:

$$R^2 = m \sum_{k=1}^m \text{tr}(\tilde{\mathbf{X}}_k - \mathbf{Z})^T (\tilde{\mathbf{X}}_k - \mathbf{Z}). \quad (8)$$

Aby wykazać, że funkcje (7) i (8) są równoważne, dokonajmy następującego przekształcenia (por. [1, s. 364]):

$$\begin{aligned} R^2 &= \sum_{k < l}^m \text{tr}(\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l)^T (\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l) = \frac{1}{2} \sum_{k=1}^m \sum_{l=1}^m \text{tr}(\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l)^T (\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l) = \\ &= \frac{1}{2} \sum_{k=1}^m \sum_{l=1}^m \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k + \frac{1}{2} \sum_{k=1}^m \sum_{l=1}^m \tilde{\mathbf{X}}_l^T \tilde{\mathbf{X}}_l - \sum_{k=1}^m \sum_{l=1}^m \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_l. \end{aligned} \quad (9)$$

Ponieważ suma dwóch pierwszych składników wynosi $m \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k$, nato-

miast trzeci składnik jest równy $\sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \sum_{l=1}^m \tilde{\mathbf{X}}_l$, zatem mamy

$$\begin{aligned}
R^2 &= m \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \sum_{l=1}^m \tilde{\mathbf{X}}_l = m \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - m \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \left(m^{-1} \sum_{l=1}^m \tilde{\mathbf{X}}_l \right) = \\
&= m \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z}.
\end{aligned}$$

Dokonując dalszych przekształceń, otrzymujemy

$$\begin{aligned}
R^2 &= m \left[\sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - 2 \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z} + \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z} \right] = \\
&= m \left[\sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - 2 \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z} + m \left(m^{-1} \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k \right)^T \mathbf{Z} \right] = \\
&= m \left[\sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - 2 \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z} + m \left(\text{tr} \mathbf{Z}^T \mathbf{Z} \right) \right] = \\
&= m \left[\sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k - 2 \sum_{k=1}^m \text{tr} \tilde{\mathbf{X}}_k^T \mathbf{Z} + \sum_{k=1}^m \text{tr} \mathbf{Z}^T \mathbf{Z} \right] = m \sum_{k=1}^m \text{tr} (\tilde{\mathbf{X}}_k - \mathbf{Z})^T (\tilde{\mathbf{X}}_k - \mathbf{Z}).
\end{aligned}$$

Równanie (7) wskazuje, że uogólniona analiza Procrustes może być dokonywana przez kolejne dopasowywanie konfiguracji $\tilde{\mathbf{X}}_k$ ($k = 1, 2, \dots, m$) do uśrednionej konfiguracji $\mathbf{Z}$, przy jednoczesnym uaktualnianiu $\mathbf{Z}$ w kolejnych cyklach iteracyjnych. Proces dopasowywania $\tilde{\mathbf{X}}_k$ jest analogiczny do procesu dla dwóch wyników skalowania wielowymiarowego.

W praktyce możemy spotkać się sytuacją występowania w niektórych macierzach $\mathbf{X}_k$ brakujących danych. Mogą to być zarówno brakujące kolumny (współrzędne punktów), jak i brakujące wiersze (punkty reprezentujące badane obiekty). W pierwszym przypadku, przed przystąpieniem do analizy Procrustes, brakujące kolumny należy zastąpić kolumnami złożonymi z samych zer. Gdy w macierzach $\mathbf{X}_k$ występują brakujące wiersze, funkcja dopasowania wyrażona jest równaniem (por. [3, s. 136]):

$$R^2 = \sum_{k=1}^m \text{tr} \left[\mathbf{M}_k (\tilde{\mathbf{X}}_k - \mathbf{Y})^T \mathbf{M}_k (\tilde{\mathbf{X}}_k - \mathbf{Y}) \right], \quad (10)$$

gdzie $\mathbf{M}_k$ jest diagonalną macierzą o wymiarach $n \times n$, przy czym i -ty diagonalny element jest równy jedności, gdy i -ty wiersz macierzy $\mathbf{X}_k$ występuje, a zero, gdy

go nie ma, natomiast $\mathbf{Y} = \left(\sum_{k=1}^m \mathbf{M}_k \right)^{-1} \sum_{k=1}^m \mathbf{M}_k \tilde{\mathbf{X}}_k$.

4. Algorytm uogólnionej analizy Procrustes na przykładzie modelu PINDIS

Model PINDIS (*Procrustean INdividual Differences Scaling*) należy do grupy modeli różnic indywidualnych (zob. np. [5]), w których dokonuje się porównania różnych wyników skalowania wielowymiarowego. W modelu PINDIS konfiguracje $\mathbf{X}_k$ porównywane są z wykorzystaniem analizy Procrustes.

W pierwszym etapie dokonuje się translacji i wyskalowania $\mathbf{X}_k$ ($k = 1, 2, \dots, m$) tak, aby ich środki ciężkości znajdowały się w początku układu współrzędnych, a średni kwadrat odległości punktów od początku układu współrzędnych był równy jedności, tzn. $\text{tr}(\mathbf{X}_k^T \mathbf{X}_k) = 1$.

W następnym kroku szacowana jest konfiguracja $\mathbf{Z}$. W tym celu dopasowuje się (wykorzystując analizę Procrustes dla dwóch wyników skalowania wielowymiarowego) $\mathbf{X}_2$ do $\mathbf{X}_1$, a następnie wyznacza macierz $\mathbf{Z}$ jako $\mathbf{Z} = \frac{1}{2}(\mathbf{X}_1 + \tilde{\mathbf{X}}_2)$. Po dopasowaniu $\mathbf{X}_3$ do oszacowanej wcześniej $\mathbf{Z}$, obliczana jest nowa wartość $\mathbf{Z} = \frac{1}{3}(\mathbf{X}_1 + \tilde{\mathbf{X}}_2 + \tilde{\mathbf{X}}_3)$. Proces ten powtarzany jest tak długo, aż wykorzystane zostaną wszystkie konfiguracje $\mathbf{X}_k$.

W następnym etapie, po ponownym dopasowaniu wszystkich konfiguracji $\tilde{\mathbf{X}}_k$ do oszacowanej $\mathbf{Z}$, oblicza się miarę dopasowania:

$$h = \frac{1}{m} \sum_{k=1}^m \left(1 - R_k^2(\tilde{\mathbf{X}}_k, \mathbf{Z}) \right)^{1/2}, \quad (11)$$

gdzie $R_k^2(\tilde{\mathbf{X}}_k, \mathbf{Z})$ jest miarą dopasowania $\tilde{\mathbf{X}}_k$ do $\mathbf{Z}$, czyli:

$$R_k^2(\tilde{\mathbf{X}}_k, \mathbf{Z}) = \frac{\left\{ \text{tr}(\tilde{\mathbf{X}}_k^T \mathbf{Z} \mathbf{Z}^T \tilde{\mathbf{X}}_k) \right\}^{1/2}}{\text{tr}(\tilde{\mathbf{X}}_k^T \tilde{\mathbf{X}}_k) \text{tr}(\mathbf{Z}^T \mathbf{Z})}. \quad (12)$$

Po wyznaczeniu nowej wartości $\mathbf{Z}$, jako średniej z otrzymanych w poprzednim etapie $\tilde{\mathbf{X}}_k$, wszystkie $\tilde{\mathbf{X}}_k$ ponownie dopasowywane są kolejno do $\mathbf{Z}$, a następnie oblicza się miarę dopasowania h . Proces powtarzany jest tak długo, aż różnica między h wyznaczonymi w kolejnych cyklach iteracyjnych przestanie ulegać zmianie lub będzie mniejsza od ustalonej stałej.

4. Podsumowanie

Niezwykle ważnym elementem skalowania wielowymiarowego jest interpretacja wyników, polegająca zarówno na określeniu treści wymiarów uzyskanej przestrzeni, jak i na identyfikacji zależności występujących między badanymi obiektami. Ponieważ nawet dla jednego badania możliwe jest uzyskanie na mapach percepcyjnych różnych konfiguracji punktów reprezentujących obiekty, przeto uzasadnione jest stosowanie procedur, dzięki którym różne wyniki końcowe doprowadzane są do porównywalności.

Znaczenie uogólnionej analizy Procrustes widoczne jest głównie wtedy, gdy w badaniu relacji zachodzących między obiektami korzysta się z opinii wielu respondentów. Jej zastosowanie pozwala otrzymać zarówno grupową konfigurację (*centroid configuration*), jak i porównywalne indywidualne konfiguracje, umożliwiające badanie relacji między obiektami w ramach kryteriów ocen stosowanych przez różnych respondentów.

Literatura

- [1] Borg I., Groenen P., *Modern Multidimensional Scaling. Theory and Applications*, Springer-Verlag, New York 1997.
- [2] Commandeur J.J.F., *Matching Configurations*, DSWO Press, Leiden 1991.
- [3] Cox T.F., Cox M.A.A., *Multidimensional Scaling*, (second edition), Chapman and Hall, London 2001.
- [4] Zaborski A., *Modele różnic indywidualnych w skalowaniu wielowymiarowym*, [w:] *Materiały z XXXIII Konferencji Statystyków, Ekonometryków i Matematyków Polski Południowej*, AE, Wrocław 1997, s. 149-159.
- [5] Zaborski A., *Porównywanie wyników skalowania wielowymiarowego metodą Procrustes*, *Ekonometria 14. Prace Naukowe Akademii Ekonomicznej nr 1021*, AE, Wrocław 2004, s. 50-59.

UTILIZING PROCRUSTES ANALYSIS FOR MANY RESULTS OF MULTIDIMENSIONAL SCALING

Summary

Procrustes analysis is a technique of matching one configuration of points obtained from multidimensional scaling analysis to another configuration obtained from a different analysis. The article presents generalized Procrustes analysis which transforms many individual configurations at the same time so that each configuration matches all others as closely as possible.

Dr Andrzej Zaborski jest pracownikiem Katedry Ekonometrii i Informatyki w Akademii Ekonomicznej we Wrocławiu.