Warianty modeli MDS

Joanna Wisła

Wrocław, 2020

1 Warianty modeli MDS

MDS to rodzina powiązanych modeli. Wszystkie z nich mierzą bliskości pomiędzy punktami. W zależności od danych oraz od tego, w jaki sposób chcemy zmierzyć bliskości punktów możemy zastosować inny model MDS. Dzięki temu możemy wykorzystywać MDS do różnych celów, dopasowując model do danych jakie posiadamy.

1.1 Klasy modeli MDS

Możemy dokonać podziału modeli MDS np. ze względu na poziom skali. Najpopularniejszym jest model porządkowy (inaczej nazywany tez niemetrycznym). Model ten opiera się na założeniu, że wartości umieszczone są na skali porządkowej. Zatem jedynymi ważnymi i przydatnymi informacjami są rangi. W porządkowym MDS porządek rangowy odległości powinien odpowiadać porządkowi rangi danych, skąd w modelu tym funkcja:

$$f : p_{ij} \rightarrow d_{ij}(X), \tag{1}$$

która jest monotoniczna, czyli

$$f : p_{ij} < p_{kl} \rightarrow d_{ij}(X) \leqslant d_{kl}(X), \tag{2}$$

Dane brakujące w tym modelu są pomijane. Istotną rolę natomiast odgrywają tzw. WIĘZY (równe wartości danych). W większości modeli porządkowych więzy są łamane - co oznacza, że równe odległosci nie muszą przechodzić na równe odleglosci. Takie podejście nosi nazwę **pierwotnego** podejścia do więzów. Natomiast **wtórne** podejście do więzów ("utrzymaj więzy powiązane") prowadzi do dodatkowego wymogu modelu, a mianowicie:

$$f : p_{ij} = p_{kl} \rightarrow d_{ij}(X) = d_{kl}(X), \tag{3}$$

Pierwotne podejście do więzów ma zazwyczaj większe znaczenie pod względem danych.
Rozważmy badanie, gdzie ankieterzy oceniają podobieństwo 66 par państw w skali od 1 do 9. Ponieważ skala jest 9-punktowa oczywiste, że w 66 porównaniach wartości powtórzą się - automatycznie zatem zdarzą się te same wartości dla niektórych par. Jest to niezależne od tego czy ankietowany rzeczywiście odczuwa, że kraje są jednakowo do siebie podobne. Co więcej, w tym przypadku żaden z badanych nie dokonałby wiarygodnej oceny gdyby skala była 66-punktowa. Każda ocena jest bowiem mniej lub bardziej rozmyta, zatem równe ratingi nie powinny być interpretowane zbyt uważnie.

1.2 MDS metryczne

Kolejna klasa modeli MDS sięga lat 50tych (1952, Torgerson). W tych modelach wymaga się, aby f (funkcja mapowania) była pewną funkcją analityczną (zwykle monotoniczną) a nie tylko "pewną" funkcją monotoniczną, jak było to w przypadku MDS porządkowym. Dzięki takiemu podejściu łatwiej w dalszych etapach rozwijać matematyczne właściwości modeli. Ponadto metoda ta unika pewnych problemów technicznych, które posiadają modele porządkowe MDS - w szczególności zdegenerowanych rozwiązań. Wadą są większe wymagania dotyczące danych. Trudność stanowi również dopasowanie

rozwiązania do danych, ponieważ trudno jest reprezentować dane w bardziej restrykcyjnych modelach. Standardowym przykładem MDS metrycznego jest MDS przedziałowe, gdzie

$$f : p_{ij} \rightarrow a + b \cdot d_{ij}(X), \tag{4}$$

W MDS przedziałowym wymagane jest, aby zachować dane w sposób liniowy jeśli chodzi o odległości. Ma to sens, jedynie gdy dane są traktowane jako przedziałowe. Oznacza to, że zakładamy, że żadne istotne informacje o danych nie zostaną utracone, gdy przekształcimy je w powyższy sposób (oprócz, oczywiście b=0). Znaczące są zatem te stwierdzenia dotyczące danych, które po takim przekształceniu pozostaną niezmienne. Inne relacje (jak np. stosunek pomiędzy wartościami danych) nie mogą mieć znaczenia.

1.3 Odległość euklidesowa i inne

Innym kryterium klasyfikacji modeli MDS jest wybór konkretnej funkcji odległości. W psychologii popularna w modelowaniu jest rodzina metryk Minkowskiego (np. badaniu oceny różnych bodźców w rożnych warunkach). Obecnie niemal zawsze stosowana jest metryka euklidesowa, gdyż odpowiada ona naturalnemu pojęciu odległosci miedzy punktami "w linii prostej". Jednak odległości euklidesowe, jak wszystkie odległości Minkowskiego wymagają płaskiej geometrii. W szczególnych przypadkach może być zatem przydatne konstruowanie reprezentacji MDS w zakrzywionych przestrzeniach. Przykładem może być odległość na kuli. Tu najkrótsza odległość miedzy punktami traktowana jest jako długość sznurka rozpiętego nad powierzchnia kuli między tymi punktami. Takie zakrzywione geometrie czasem mogą być przydatne (np. w psychofizyce), ale nie są stosowane w ogólnych sytuacjach analizy danych.

1.4 Symetria i asymetria

Odległości są zawsze symetryczne, tzn. zawsze $d_{ij} = d_{ji}$. Dlatego obiekty, które nie są symetryczne sprawiają spore problemy w modelach MDS. Dopóki asymetrie są po prostu oparte na błędach, nie powstaje rzeczywisty problem, gdyż MDS może je wygładzić czy wyeliminować np. uśredniając. Czasem jednak zdarza się, że asymetrie niosą za sobą ważne i wiarygodne informacje.

Przykładem takich informacji mogą być tabele eksportu-importu pomiędzy państwami (X importuje więcej z Y). Inny przykład to liczba polubień w mediach społecznościowych - osoba A lubi większość zdjęć osoby B, natomiast na odwrót niekoniecznie.

Aby w MDS radzić sobie z wielkościami asymetrycznymi wprowadzono tzw. **model dryfu**. Model ten wymaga od użytkownika utworzenia dwóch macierzy z macierzy P. Pierwszą do utworzenia jest składowa symetryczna $S = (P + P')/2$. Druga to składowa skośnie-symetryczna $A = P - S$ z elementami $a_{ij} = p_{ij} - s_{ij}$. Pierwszą składową w sposób standardowy przekształca (skaluje) się przez model MDS. Natomiast drugą składową wykorzystuje się w taki sposób, że do każdego punktu i dołączona zostaje strzałka, która wskazuje na punkt j lub w kierunku przeciwnym do j, w zależności od znaku asymetrii. Długość tej strzały to $k \cdot \mid aij \mid$, gdzie k to pewien współczynnik.

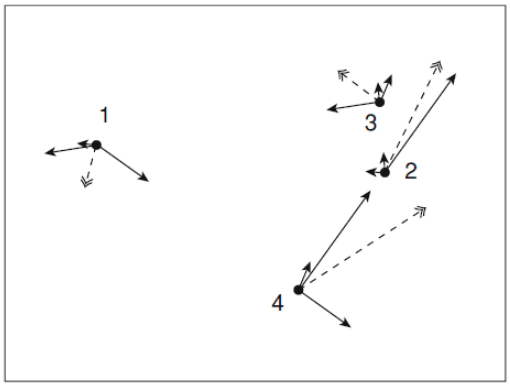


Fig. 5.2 Vector field over an MDS solution; *dashed arrows* are resultants

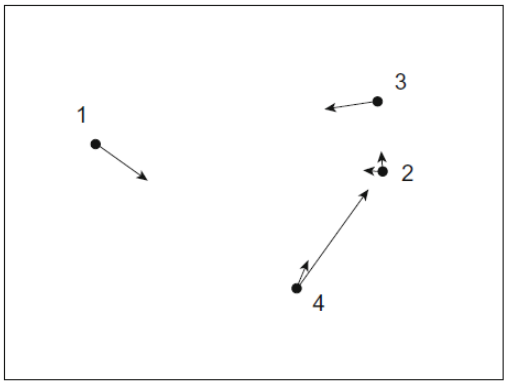
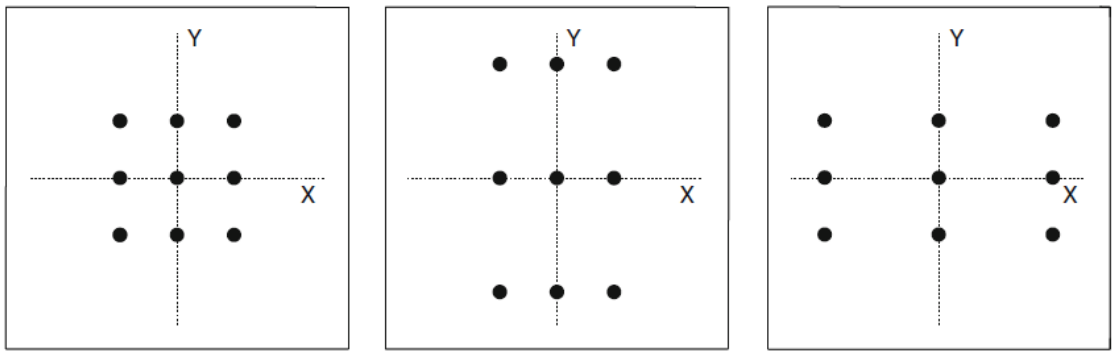


Fig. 5.3 Vector field reduced to positive asymmetries

Na podstawie pewnej macierzy danych P zastosowano model dryfu. Na rysunku *Fig. 5.3* pole wektorowe z rysunku *Fig. 5.2* zostało uproszczone, pokazane są na nim tylko te strzałki, które są dodatnio skierowane od i do j. Widać, że pomiędzy obiektami 2 i 3 istnieje silny związek (gdyż są blisko siebie) oraz ich relacja jest symetryczna (krótki wektor dryfu). Natomiast dla obiektów 4 i 2 związek jest słabszy (odległość między punktami jest większa), ponadto rozkłada się to asymetrycznie (długi wektor dryfu).
Nie uzyskano do tej pory wystarczająco prostych programów komputerowych, które w łatwy sposób eksperymetowałyby z przedstawianiem asymetrii za pomocą wektora dryfu w różnych konfiguracjach. Mimo tego, że wyniki są często skomplikowane i niewygodne w użytkowaniu to ich rezultaty są warte wysiłku, gdyż mogą ujawnić asymetrie (nad symetryczną struktura bazowa) - a te są trudne do wykrycia w macierzy danych.

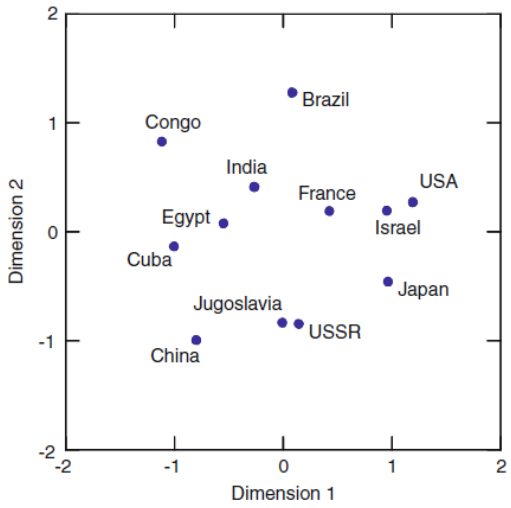
1.5 Modelowanie indywidualnych różnic w MDS



Jednym z popularnych wariantów MDS jest model ważenia wymiarowego, inaczej nazywamy modelem Indscala. Powyższy rysunek przedstawia schematyczną idee tego modelu. Powyższe wykresy *Individual space 1* i *Individual space 2* przedstawiają psychologiczne przestrzenie dwóch osób. Przestrzenie te są różne, jednak można je wygenerować z jednej wspólnej przestrzeni (znajdującej się na wykresie *Common space*) za pomocą odpowiednich wag o wymiarach X i Y. W przeciwieństwie do wariantów MDS, wymiary w modelu Indscal są w ogólności stałe.
Przeanalizujmy wyniki badania, którego celem jest poznanie atrybutów używanych przez ludzi podczas oceniania podobieństwa różnych krajów. Osiemnaście respondentów oceniło podobieństwo dla każdej pary dwunastu krajów. Skrajnie odmienne państwa oznaczano 1, a bardzo podobne 9. Zaobserwowane wartości uśredniono i przedstawiono w poniższej tabeli.

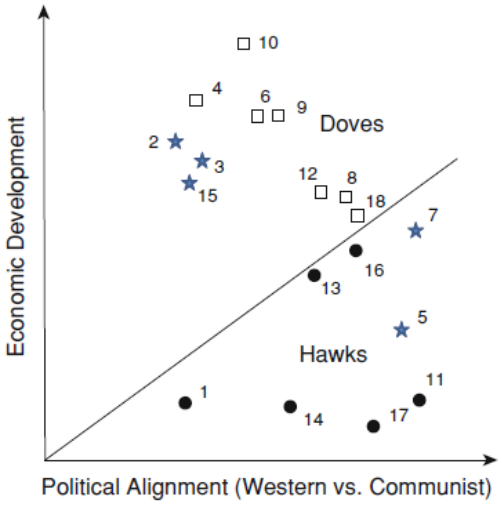
Mean similarity ratings for 12 countries (Wish 1971)

Country		1	2	3	4	5	6	7	8	9	10	11	
Brazil	1	–											
Congo	2	4.83	–										
Cuba	3	5.28	4.56	–									
Egypt	4	3.44	5.00	5.17	–								
France	5	4.72	4.00	4.11	4.78	–							
India	6	4.50	4.83	4.00	5.83	3.44	–						
Israel	7	3.83	3.33	3.61	4.67	4.00	4.11	–					
Japan	8	3.50	3.39	2.94	3.83	4.22	4.50	4.83	–				
China	9	2.39	4.00	5.50	4.39	3.67	4.11	3.00	4.17	–			
USSR	10	3.06	3.39	5.44	4.39	5.06	4.50	4.17	4.61	5.72	–		
USA	11	5.39	2.39	3.17	3.33	5.94	4.28	5.94	6.06	2.56	5.00	–	
Jugoslavia	12	3.17	3.50	5.11	4.28	4.72	4.00	4.44	4.28	5.06	6.67	3.56	–



Kruskal i Wish użyli modelu Indscala do skalowania surowych danych z tego badania. Przeanalizowali oni postawione przez respondentów oceny za pomocą modelu Indscala i otrzymali rozwiązanie wspólnej przestrzeni, która jest porównywalna do tej przedstawionej na rysunku (po lewej stronie). Przedstrzeń ta różni się jednak tym, że jest ona obrócona o około 45^0.

Na poniższym wykresie przedstawiono punkty (masy) obliczone przez model. Wykres ten w modelu Indscala nazywa się *przestrzenią przedmiotu*. Interpretacja punktów jest następująca, gdy na przykład weźmiemy osobe 11 to mówimy, że znacznie przewyższa ona wagę wymiaru X. Oznacza to, że osoba ta poświęca stosunkowo niewiele uwagi rozwojowi gospodarczemu w ocenach podobieństwa krajów, jednak zwraca dużo uwagi na polityczne dopasowanie krajów. Z kolei dla osoba 4 interpretacja ta jest odwrotna, a więc znacznie przewyższa ona wagę wymiaru Y, czyli osoba ta skupia się bardziej na rozwoju gospodarczym państw niż na politycznym ich dopasowaniu. W tym przypadku interpretacja została uzupełniona również dodatkowymi danymi dotyczącymi respondentów. Na podstawie pytania o wojnę w Wietnamie zostali oni podzieleni na trzy grupy: jastrzębi (punkty), gołębi (kwadraty) i umiarkowanych (gwiazdki). Grupy te pojawiają się w przestrzeni przedmiotowej w spodziewanych regionach wag. Odległość punktu i od początku przestrzeni przedmiotowej Indscala odpowiada stopniowi dopasowania modelu Indscal dla i-tej osoby. Wykres pokazuje zatem na przykład, że dane dotyczące osoby 1 są stosunkowo słabo wyjaśnione, zaś osób - 10, 7 lub 11 są wyjaśnione bardzo dobrze.



Procedura Indscala może być łatwo zinterpretowana. Jednym problemem jest to, czy dopasowanie modelu byłoby wyraźnie gorsze, gdyby wszystkie wagi wymiarów miały ustawioną taką samą wartość. Masy wymiarów zależą od konkretnego systemu wymiarów wspólnej przestrzeni. Z czego wynika, że nie można wywnioskować z wykresu, że osoba 11 zwraca uwagę na polityczne dostosowanie krajów 6 razy bardziej niż ich rozwój gospodarczy. Można jednak stwierdzić, że osoba 11 ma większą masę wymiaru X niż osoba 10, ponieważ ta relacja pozostaje niezmienna.

Model Indscala jest interesujący, ale w praktyce rzadko prowadzi do przekonujących, pod względem dopasowania danych, rozwiązań. Innym powodem jest to, że model ten jest często nieodpowiednim modelem psychologicznym do oceny podobieństwa. Badania pokazały, że różne osoby mogą generować bardzo różne "wymiary" nawet najprostszych bodźców, czyli takich, których wymiary wydają się oczywiste. Ponadto mogą wykorzystywać różne funkcje odległościowe, niespełniające nawet podstawowych aksjomatów odległości. Jednak formułowanie i studiowanie modela Indscala znacznie przyczyniło się do dzisiejszego, bardziej wyrafinowanego rozumienia psychologicznych sądów o podobieństwie.

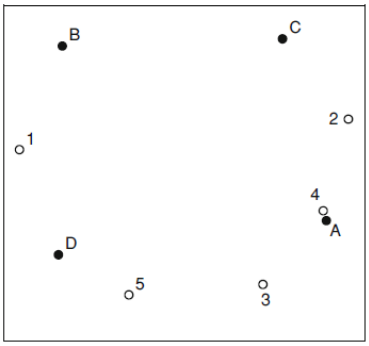
1.6 Model rozwijania (Unfolding)

Kolejnym popularnym wariantem MDS jest model rozwijania (Unfolding). Stosuje się go do danych dominujących, a więc dla takich, które wyrażają stopień, w jakim i dominuję nad j w pewnym sensie (np. głosowałbym raczej na A niż na B) lub dla danych dotyczących preferencji M osób nad n obiektami.

W tabelce pokazano dane dotyczące pięciu osób (oznaczonych liczbami) i czterech partii politycznych (oznaczonych literami). Liczby przedstawiają wartości preferencji (4 oznacza dużą skłonność osoby do danej partii, a 1 bardzo niską). Osoba 3 najbardziej preferuje partę A, następnie D oraz C, a na samym końcu najmniej preferuje partię B. Taką macierz danych można rozumieć jako szczególny przypadek macierzy bliskości, w której dane wyrażają stopień bliskości danej osoby do danej partii.

		Parties				Persons				
		A	B	C	D	1	2	3	4	5
Parties	A	-	-	-	-	1	4	4	4	3
	B	-	-	-	-	3	2	1	1	2
	C	-	-	-	-	2	3	2	3	1
	D	-	-	-	-	4	1	3	2	4
Persons	1	1	3	2	4	-	-	-	-	-
	2	4	2	3	1	-	-	-	-	-
	3	4	1	2	3	-	-	-	-	-
	4	4	1	3	2	-	-	-	-	-
	5	3	2	1	4	-	-	-	-	-

Używając zwykłego MDS do skalowania tych danych, otrzymujemy 9 punktów, 5 punktów oznaczających osoby i 4 partie, jak pokazano na rysunku poniżej. Punkty reprezentujące osoby nazywane są **idealnymi punktami** w modelu rozwijania (unfloding), ponieważ są to punkty maksymalnych preferencji w przestrzeni: im bliżej partii do idealnego punktu danej osoby, tym silniejsza jest skłonność tej osoby do tej partii. Odległość między idealnym punktem, a punktem reprezentującym partie określają skłonność danej osoby do partii. Dlatego, składanie (jak składanie parasola lub podnoszenie chusteczki w danym punkcie) rozwijanej (unfloding) płaszczyzny w idealnym punkcie prowadzi od oceny preferencji danej osoby dla różnych partii. Oczywiście, złożenie płaszczyzny w innym punkcie generuje inne wyniki preferencji.



Jako psychologiczny model preferencji, model rozwijający (unfloding) opiera się na założeniu, że wszystkie osoby mają takie samo postrzeganie obiektów. Mogą różnić się tylko tym, co uważają za idealne. Z geometrycznego punktu widzenia model rozwijania (unfloding) może z łatwością doprowadzić do niestabilnych rozwiązań. Dlatego, że opiera się on na danych, które ograniczają tylko podzbiór odległości, a mianowicie odległości między idealnymi punktami, a punktami obiektowymi, ale nie odległości między idealnymi punktami oraz również nie odległości między punktami obiektu. Wybierając model rozwijający należy również rozważyć, czy naprawdę chcemy założyć, że dane są porównywalne we wszystkich wierszach. W naszym przykładzie można wątpić, że preferencyjna wartość "4" osoby 1 jest prawdziwie równa "4" osoby 2. Jeśli nie chcemy założyć, że jednakowe wartości danych mają to samo znaczenie, to odległość od idealnego punktu osoby 1 do punktu D nie powinna mieć takiej samej wartość jak odległość od idealnego punktu osoby 2 do punktu A.

1.7 Podsumowanie

MDS to rodzina wielu różnych wariantów modeli. Różnią się one sposobem, w jaki mapują bliskości punktów i funkcjami odległości, które wykorzystują. Różne funkcje mapowania zachowują pewne właściwości danych, na przykład szeregi danych w porządkowych MDS-ach, względne różnice dowolnych dwóch wartości danych w przedziałowych MDS-ach lub stosunek danych w proporcjonalnych MDS-ach. Najczęściej wybiera się odległości euklidesowe, jednak również odległości między miastami a wskaźnikami dominacji są wykorzystywane w modelowaniu psychologicznym. Niektóre modele MDS pozwalają na używanie wielu bliskości. Asymetryczne bliskości mogą być obsługiwane przez model dryftu. Popularnym modelem MDS jest model Indscal, który przedstawia jedną wspólną przestrzeń danych i poprzez punkty na wykresie wagi wymiarów tych danych. Kolejnym modelem MDS jest model rozwijający (unfloding). Korzysta on z danych dotyczących preferencji, reprezentujących M osób przez M punktów idealnych we wspólnej przestrzeni wraz z punktami dla różnych obiektów wyboru.